Phonetic recalibration in audiovisual speech

Baart, M.

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

PHONETIC RECALIBRATION IN AUDIOVISUAL SPEECH

Martijn Baart

Phonetic recalibration in audiovisual speech

ISBN: 978-90-5335-511-4

Printed: Ridderprint BV, Ridderkerk
Cover design: Paul Baart

Phonetic recalibration in audiovisual speech

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof.dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 2 maart 2012 om 14.15 uur door

Martijn Baart,

geboren op 24 juli 1983 te Hulst.

Promotor
Prof. dr. J. H. M. Vroomen

Promotiecommissie
Prof. dr. A. G. Samuel
Prof. dr. J. M. McQueen
Dr. J. Tuomainen
Dr. J. J. Stekelenburg

Contents

Chapter 1

Introduction [1]

[1]Adapted from:

Vroomen, J., & Baart, M. (2011). Phonetic recalibration in audiovisual speech. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 363-379). Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.

1.1 - General introduction

In the literature on cross-modal perception, there are two important findings that most researchers in that area will know about. However, only few have ever made a connection between the two. The first is that perceiving speech is not solely an auditory, but a multi-sensory phenomenon as seeing a speaker's face can help decoding the spoken message (Erber, 1974; Sumby & Pollack, 1954). The most famous experimental demonstration of the multisensory nature of speech is the so-called McGurk-illusion: when perceivers are presented an auditory syllable /ba/ dubbed onto a face articulating /ga/, they report 'hearing' /da/ (McGurk & MacDonald, 1976). The second finding goes back more than 100 years (Stratton, 1896). Stratton did experiments with goggles and prisms that radically changed his visual field, thereby creating a conflict between vision and proprioception. What he experienced is that after wearing prisms for a couple of days, he adapted to the upside-down visual world and he learned to move along in it quite well. According to Stratton, the visual world had changed as it sometimes appeared to him as if it was 'right side up', although later, others like Held (1965) argued that it rather was the sensor-motor system that was adapted.

What these two seemingly different phenomena have in common is that in both cases, an artificial conflict between the senses is created about an event that should yield congruent data under normal circumstances. Thus, in the McGurk-illusion, there is a conflict between the auditory system that hears the syllable /ba/ and the visual system that sees the face of a speaker saying /ga/; in the prism case there is a conflict between proprioception that may feel the hand going upwards and the visual system that sees the same hand going downwards. A couple of years ago, the commonality between these two phenomena led Bertelson, Vroomen and de Gelder (2003) to the question whether one might also observe long-term adaptation effects with audiovisual speech as reported by Stratton for prism adaptation. To be more specific, presumably nobody had ever examined whether auditory speech perception would adapt as a consequence of exposure to the audiovisual conflict present in McGurk-stimuli. Actually, this was rather surprising given that the original paper by McGurk and MacDonald is one of the most highly-cited papers in this research area.

Admittedly, though, on first sight it may look as a somewhat exotic enterprise to examine whether listeners adapt to speech sounds induced by exposure to an audiovisual conflict. After all, why would adaptation to a video of an artificially dubbed speaker be of importance? Experimental psychologists should rather spend their time on fundamental aspects of perception and cognition that remain constant across individuals, cultures, and time, and not on matters that are flexible and adjustable. And indeed, the dominant approach in speech research did just that by focusing on the

information available in the speech signal, the idea being that there must be acoustic invariants in the signal that are extracted during perception. On second thought, though, it has turned out to be extremely difficult to find a set of acoustic invariant parameters that work for all contexts, cultures, and speakers, and the question that Bertelson et al. (2003) addressed might support an alternative view: Rather than searching for acoustic invariants, it might be equally fruitful to examine whether and how listeners adjust their phoneme boundaries so as to accommodate the variation they hear. And indeed, in 2003, Bertelson et al. reported that phonetic recalibration induced by McGurk-like stimuli can be observed. The authors termed the phenomenon 'recalibration' in analogy with the much better known 'spatial recalibration', as they considered it a re-adjustment or a fine-tuning of an already existing phonetic representation. In the same year, and in complete independence, Norris, McQueen, and Cutler (Norris, McQueen, & Cutler, 2003) reported a very similar phenomenon they named 'perceptual learning in speech'. The basic procedure in both studies was very similar: Listeners were presented with a phonetically ambiguous speech sound and another source of contextual information that disambiguated that sound. Bertelson et al. (2003) presented listeners a sound halfway between /b/ and /d/ with the video of a synchronized face that articulated /b/ or /d/ (in short, lipread information) as context, while in Norris et al. (2003), an ambiguous /s/-/f/ sound was heard embedded in the context of an f- or s-biasing word (e.g., 'witlo-s/f' was an f-biasing context because 'witlof' is a word in Dutch meaning 'chicory', but 'witlos' is not a Dutch word). Recalibration (or perceptual learning) was subsequently measured in an auditory-only identification test in which participants identified members of a speech continuum. Recalibration manifested itself as a shift in phonetic categorization toward the contextually-defined speech environment. Listeners thus increased their report of sounds consistent with the context they had received before, so more /b/ responses after exposure to lipread /b/ rather than lipread /d/, and more /f/ responses after exposure to /f/-biasing words rather than /s/-biasing words. Presumably, this shift reflected an adjustment of the phoneme boundary that had helped listeners to understand speech better in the prevailing input environment. Following these seminal reports, there have been a number of studies that examined phonetic recalibration in more detail (Baart & Vroomen, 2010a, 2010b; Cutler, McQueen, Butterfield, & Norris, 2008; Eisner & McQueen, 2005, 2006; Jesse & McQueen, 2011; Kraljic & Samuel, 2005, 2006; Kraljic, Samuel, & Brennan, 2008; McQueen, Cutler, & Norris, 2006; McQueen, Jesse, & Norris, 2009; McQueen, Norris, & Cutler, 2006; Sjerps & McQueen, 2010; van Linden, Stekelenburg, Tuomainen, & Vroomen, 2007; van Linden & Vroomen, 2007, 2008; Vroomen & Baart, 2009a, 2009b; Vroomen, van Linden, de Gelder, & Bertelson, 2007; Vroomen, van Linden, Keetels, de Gelder, & Bertelson,

2004). The following sections contain an overview of the relevant literature and a theoretical framework.

1.2 - A short historic background on audiovisual speech aftereffects

Audiovisual speech has been extensively studied in recent decades ever since seminal reports that lipread information is of help in noisy environments (Sumby & Pollack, 1954) and, given appropriate dubbings, can change the auditory percept (McGurk & MacDonald, 1976). More recently, audiovisual speech has served in functional magnetic resonance imaging (fMRI)-studies as an ideal stimulus for studying the neural substrates of multisensory integration (Calvert & Campbell, 2003). Surprisingly, though, until 2003 there were only three studies that had focused on auditory aftereffects as a consequence of exposure to audiovisual speech.

Roberts and Summerfield (1981) were the first to study aftereffects of audiovisual speech, though they were not searching for recalibration, but 'selective speech adaptation', which is basically a contrastive effect. The main question of their study was whether selective speech adaptation takes place at a phonetic level of processing, as originally proposed by Eimas and Corbit (1973), or at a more peripheral acoustic level. Selective speech adaptation differs from recalibration in that it does not depend on an (intersensory) conflict, but rather on the repeated presentation of an acoustically non-ambiguous sound that reduces report of sounds similar to the repeating one. For example, hearing /ba/ many times reduces subsequent report of /ba/ on a /ba/-/da/ test continuum. Eimas and Corbit (1973) argued that selective speech adaptation reflects the neural fatigue of hypothetical 'linguistic feature detectors', but this viewpoint was not left unchallenged by others claiming that it reflects a mere shift in criterion (Diehl, 1981; Diehl, Elman, & McCusker, 1978; Diehl, Lang, & Parker, 1980) or a combination of both (Samuel, 1986) or possibly, that even more qualitatively different levels of analyses are involved (Samuel & Kat, 1996). Still others (Sawusch, 1977) showed that the size of selective speech adaptation depends upon the degree of spectral overlap between the adapter and test sound, and that most - though not all of the effect - is acoustic rather than phonetic.

Roberts and Summerfield (1981) found a clever way to disentangle the acoustic from the phonetic contribution using McGurk-like stimuli. They dubbed a canonical auditory /b/ (a 'good' acoustic example) onto the video of lipread /b/ to create an audiovisual congruent adapter and also dubbed the auditory /b/ onto a lipread /g/ to create a compound stimulus intended to be perceived as /d/. Results showed that repeated exposure to the congruent audiovisual adapter induced similar contrastive aftereffects on a /b/-/d/ test continuum (i.e.,, fewer /b/ responses) as the incongruent

adapter AbVg, even though the two adapters were perceived differently. This led the authors to conclude that selective speech adaptation mainly depends on the acoustic quality of the stimulus, and not the perceived or lipread one.

Saldaña and Rosenblum (1994) and Shigeno (2002) later replicated these result with different adapters. Saldaña and Rosenblum compared auditory-only adapters with audiovisual ones (auditory /b/ paired with visual /v/, a compound stimulus perceived mostly as /v/), and found, as in Roberts and Summerfield, that the two adapters again behaved similarly, as in both cases fewer /b/ responses were obtained at test. Similar results were also found by Shigeno (2002) using AbVg as adapter, demonstrating that selective speech adaptation depends, to a large extent, on repeated exposure to non-ambiguous sounds.

1.3 - The seminal study on lipread-induced recalibration

Bertelson et al. (2003) also studied aftereffects of audiovisual incongruent speech, but their focus was not on selective speech adaptation but on recalibration. Their study was inspired by previous work on aftereffects of the 'ventriloquist illusion'. In the ventriloquist illusion, the apparent location of a target sound is shifted towards a visually displaced distracter that moves or flashes in synchrony with that sound (Bermant & Welch, 1976; Bertelson & Aschersleben, 1998; Bertelson & Radeau, 1981; Klemm, 1909). Besides this immediate bias in sound localization, one can also observe aftereffects following prolonged exposure to a ventriloquized sound (Bertelson, Frissen, Vroomen, & De Gelder, 2006; Radeau & Bertelson, 1974, 1976, 1977). For the ventriloquist situation, it was known that the location of target sounds shifts towards the visual distracter seen during the preceding exposure phase. These aftereffects were similar to the ones following exposure to discordant visual and proprioceptive information – such as when the apparent location of a hand is displaced through a prism (Welch & Warren, 1986) – and they all showed that exposure to spatially conflicting inputs recalibrates processing in the respective modalities in a way that reduces the conflict.

Despite the fact that immediate biases and recalibration effects had been demonstrated for spatial conflict situations, the existing evidence was less complete for conflicts regarding audiovisual speech. Here, immediate biases were well-known (the McGurk-effect) as well as selective speech adaptation, but recalibration had not been demonstrated. Bertelson et al. (2003) hypothesized that a slight variation in the paradigm introduced by Roberts and Summerfield (1981) might nevertheless produce these effects, thus revealing recalibration. The key factor was the ambiguity of the adapter sound. Rather than using a conventional McGurk-like stimulus containing a

canonical (and incongruent) sound, Bertelson et al. (2003) used an ambiguous sound. They created a synthetic sound halfway between /aba/ and /ada/ (henceforth A? for auditory ambiguous) and dubbed it onto the corresponding video of a speaker pronouncing /aba/ or /ada/ (A?Vb and A?Vd, respectively). Participants were shortly exposed to either A?Vb or A?Vd and then tested on identification of A? and the two neighbor-tokens on the auditory continuum A?-1 and A? +1. Each exposure block contained 8 adapters (either A?Vb or A?Vd) immediately followed by 6 test trials. These exposure-test blocks were repeated many times and participants were thus biased towards both /b/ and /d/ in randomly ordered blocks (a within-subjects factor). Results showed that listeners quickly learned to label the ambiguous sound in accordance with the lipread information they were exposed to shortly before. Listeners thus gave *more* /aba/ responses after exposure to A?Vb than after exposure to A?Vd, and this was taken as the major sign of recalibration (see Figure 1a; left panel).

In a crucial control experiment, Bertelson et al. (2003) extended these findings by incorporating audiovisual congruent adapters AbVb and AdVd. These adapters were not expected to induce recalibration because there was no conflict between sound and vision. Rather, they were expected to induce selective speech adaptation due to the non-ambiguous nature of the sound. As shown in Figure 1a, right panel, these adapters indeed induced selective speech adaptation, and there were thus *less* /aba/ responses after exposure to AbVb than AdVd, an effect in the opposite direction of recalibration.

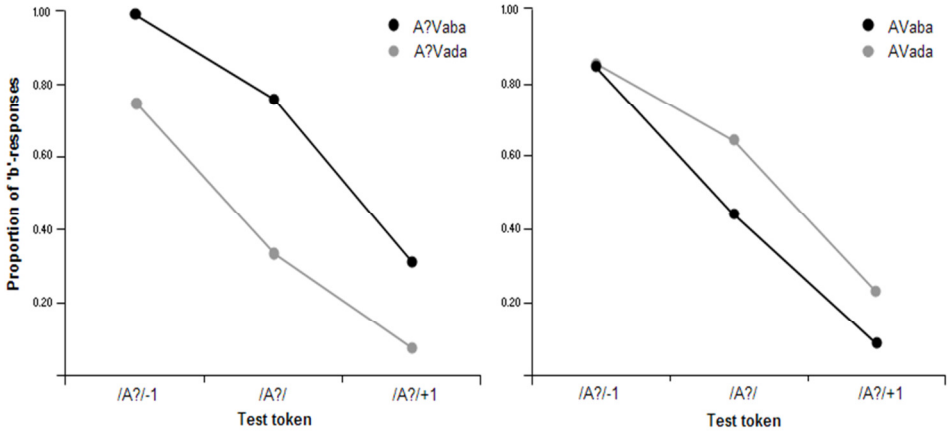

Figure 1a. *Results on the auditory tests adapted from Bertelson, Vroomen and de Gelder (2003, Exp 2). The figure shows the percentage of /aba/ responses as a function of the auditory test token. Left panel: After exposure to audiovisual adapters with ambiguous sounds, A?Vb or A?Vd, there were more responses consistent with the adapter (recalibration). Right panel: After exposure to audiovisual adapters with non-ambiguous sounds, AbVb or AdVd, there were fewer responses consistent with the adapter (selective speech adaptation).*

The attractiveness of these control stimuli was that participants could not distinguish them from the ones with an ambiguous sound that induced recalibration. This was confirmed in an identification test in which A?Vb and AbVb were perceived as /b/, and A?Vd and AdVd as /d/ on nearly 100% the trials. Moreover, even when participants were explicitly asked to discriminate AbVb from A?Vb, and AdVd from A?Vd, they performed at chance level because there was a strong immediate bias by the lipread information that captured the identity of the sound (Vroomen et al., 2004). These findings imply that the difference in aftereffects induced by adapters with ambiguous versus non-ambiguous sounds cannot be ascribed to some (unknown) explicit strategy of the listeners, because listeners simply could not know whether they were actually hearing adapters with ambiguous sounds (causing recalibration) or non-ambiguous sounds (causing selective speech adaptation). This confirms the sensory rather than strategic nature of the phenomenon.

Lipread-induced recalibration of speech was thus demonstrated and appeared to be contingent upon exposure to an ambiguous sound and another source of information that disambiguated that sound. Selective speech adaptation on the other hand, occurred in the absence of an intersensory conflict and mainly depended on repeated presentation of an acoustically clear sound. These two forms of aftereffects had been studied before in other perceptual domains, but always in isolation. Recalibration was earlier demonstrated for the ventriloquist situation and analogous intra-modal conflicts such as between different cues to visual depth (see Epstein, 1975; Wallach, 1968 for reviews), whereas contrastive aftereffects where already well-known for color, curvature (Gibson, 1933), size (Blakemore & Sutton, 1969) and motion (Anstis, 1986; Anstis, Verstraten, & Mather, 1998). Interestingly, in the ventriloquist illusion, the reverse phenomenon, namely that the perceived visual target location is shifted towards an auditory displaced distracter, has also been demonstrated (Radeau & Bertelson, 1987). In close correspondence with these bi-directional ventriloquist effects, phonetic recalibration of perceived lipread speech can also be induced by sound identity of the exposure stimuli (Baart & Vroomen, 2010a, see also Chapter 4).

1.4 - Other differences between recalibration and selective speech adaptation

After the first report, several follow-up studies examined differences in the manifestation of lipread-induced recalibration and selective speech adaptation. Besides that the two phenomena differed in the direction of their aftereffects, differences were found in their build-up, dissipation rate, and the processing mode in which they occur (i.e., ‘speech mode’ versus ‘non-speech mode’).

1.4.1 - Build-up

To examine the build-up of recalibration and selective speech adaptation, Vroomen et al. (2007) presented the four previously used audiovisual adapters (A?Vb, A?Vd, AbVb, and AdVd) in a continuous series of exposure trials, and inserted test trials after 1, 2, 4, 8, 16, 32, 64, 128, and 256 exposures. The aftereffects of adapters with ambiguous sounds (A?Vb and A?Vd) were already at ceiling after only eight exposure trials (the level of exposure used in the original study) and then, surprisingly, after 32 exposure trials fell off with prolonged exposure (128 and 256 trials). Aftereffects of adapters with non-ambiguous sounds AbVb and AdVd were again contrastive and the effect linearly increased with the (log-)number of exposure trials. The latter fitted well with the idea that selective speech adaptation reflects an accumulative process, but there was no apparent reason why a learning effect like recalibration would reverse at some point. The authors suggested that two processes might be involved here, namely, selective speech adaptation is running in parallel with recalibration and can eventually take over. Recalibration would then dominate the observed aftereffects in the early stages of exposure, whereas selective speech adaptation would become manifest later on.

Such a phenomenon was indeed observed when data of an 'early' study (i.e., one before the initial reports on phonetic recalibration) by Samuel (2001) were re-analyzed. Samuel exposed his participants to massive repeated presentations of an ambiguous /s/–/ʃ/ sound in the context of either an /s/-final word (e.g., /bronchiti?/, from bronchitis), or a /ʃ/-final one (e.g., /demoli?/, from demolish). In this situation, one might expect recalibration to take place. However, in post-tests involving identification of the ambiguous /s/–/ʃ/ sound, Samuel obtained contrastive aftereffects indicative of selective speech adaptation, so less /s/ responses after exposure to /bronchiti?/ than /demoli?/ (and thus an effect in the opposite direction later reported by Norris et al., 2003). This made him conclude that a lexically-restored phoneme produces selective speech adaptation similar to a non-ambiguous sound. In later years though, others, - including Samuel - would report recalibration effects using the same kinds of stimuli (Kraljic & Samuel, 2005; Norris et al., 2003; van Linden & Vroomen, 2007). To examine this potential conflict in more detail, the data from Samuel (2001) were re-analyzed as a function of number of exposures blocks (Vroomen, et al., 2007). Samuel's experiment consisted of 24 exposure blocks, each containing 32 adapters. Contrastive aftereffects were indeed observed for the majority of blocks following block 3, showing the reported dominant role of selective speech adaptation. Crucially though, a significant recalibration effect was obtained (so more /s/ responses after exposure to /bronchiti?/ than /demoli?/) in the first block of 32 exposure trials, which, in the overall

analyses, was swamped by selective adaptation in later blocks. Thus, the same succession of aftereffects dominated early by recalibration and later by selective adaptation was already present in Samuel's data. The same pattern may therefore occur generally during prolonged exposure to various sorts of conflict situations involving ambiguous sounds.

1.4.2 - Dissipation

A study by Vroomen et al. (2004) focused on how long recalibration and selective speech adaptation effects last over time. Participants were again exposed to A?Vb, A?Vd, AdVd, or AbVb, but rather than using multiple blocks of 8 adapters and 6 test trials in a within-subject design (as in the original study), participants were now exposed to only one of the four adapters (a between-subject factor) in three similar blocks consisting of 50 exposure trials followed by 60 test trials. The recalibration effect turned out to be very short-lived and lasted only about 6 test trials, whereas the selective speech adaptation effect was observed even after 60 test trials. The results again confirmed that the two phenomena were different from each other.

1.4.3 - Recalibration in 'speech'- versus 'non-speech' mode

The basic notion underlying recalibration is that it occurs to the extent that there is a (moderate) conflict between two information sources that refer to the same external event (for speech, a particular phoneme or gesture). Using sine-wave speech (SWS), one can manipulate whether the acoustic input is assigned to a speech sound (for short, a phoneme) or not, and thus whether recalibration occurs. In SWS, the natural richness of speech sounds is reduced and an identical sound can be perceived as speech or non-speech depending on the listener's perceptual mode (Remez, Rubin, Pisoni, & Carrell, 1981). Tuomainen, Andersen, Tiippana and Sams (2005) demonstrated that when SWS sounds are delivered in combination with lipread speech, listeners who are in speech mode show almost similar intersensory integration as when presented with natural speech (i.e.,, lipread information strongly biases phoneme identification), but listeners who do not know that the SWS tokens are derived from speech ('non-speech mode') show no, or only negligible integration. Using these audiovisual SWS stimuli, we reasoned that recalibration should only occur for listeners in speech mode (Vroomen & Baart, 2009a, see also Chapter 3). To demonstrate this, participants were first trained to distinguish the SWS tokens /omso/ and /onso/ that were the two extremes of a seven-step continuum. Participants in the speech group labelled the tokens as /omso/ or /onso/, while the non-speech group labelled the same sounds as '1' and '2'. Listeners were then shortly exposed to the adapters A?Vomso and A?Vonso (to examine recalibration),

and AomsoVomso and AonsoVonso (to examine selective speech adaptation), and then tested on the three most ambiguous SWS tokens that were identified as /omso/ or /onso/ in the speech group, and as '1'or '2' in the non-speech group. Recalibration only occurred for listeners in speech-mode, but not in non-speech mode, whereas selective speech adaptation was alike in speech- and non-speech mode. Attributing the auditory and visual signal to the same event was thus of crucial importance for recalibration, whereas selective speech adaptation did not depend on the interpretation of the signal.

1.5 - The stability of recalibration over time

As mentioned before, studies on phonetic recalibration began with a pair of seminal studies, of which one used lipread information (Bertelson, et al., 2003) and the other used lexical information (Norris, et al., 2003). Both showed in essence the same phenomenon, but the results were nevertheless strikingly different in one aspect: Whereas lipread-induced recalibration was short-lived, lexical recalibration turned out to be robust and long-lived in the majority of studies. The reasons for this difference are still not well-understood but the overview below contains the main findings and some hints on possible causes.

1.5.1 - The basic phenomenon of lexically-induced recalibration

It is well-known that in natural speech, there are, besides the acoustic and lipread input, other information sources that inform listeners about the identity of the phonemes. One of the most important ones is the listener's knowledge about the words in the language, or for short, lexical information. As an example, listeners can infer that an ambiguous sound somewhere in between /b/ and /d/ in the context of '?utter' is more likely to be /b/ rather than /d/ because 'butter' is a word in English, but not 'dutter'. There is also, as is the case for lipreading, an immediate lexical bias in phoneme identification known as the Ganong-effect (Ganong, 1980). For example, an ambiguous /g/-/k/ sound is 'heard' as /g/ when followed by 'ift' and as /k/ when followed by 'iss' because 'gift' and 'kiss' are words, but 'kift' and 'giss' are not.

The corresponding aftereffect that results from exposure to such lexically-biased phonemes was first reported by Norris et al. (2003). They exposed listeners to a sound halfway between /s/ and /f/ in the context of an f- or s-biasing word, and listeners were then tested on an /es/-/ef/ continuum. The authors observed recalibration (or in their terminology, perceptual learning) comparable to the lipreading case, so more /f/ responses after an f-biasing context, and more /s/ responses after an s-biasing context.

Later studies confirmed the original finding and additionally suggested that the effect is speaker-specific (Eisner & McQueen, 2005) or possibly, token-specific (Kraljic

& Samuel, 2006, 2007), that it generalizes to words outside the original training set (McQueen, Cutler, et al., 2006) and across syllabic positions (Jesse & McQueen, 2011), and that it arises automatically as a consequence of hearing the ambiguous pronunciations in words (McQueen, Norris, et al., 2006). Although Jesse and McQueen (2011) demonstrated that lexical recalibration can generalize to word onset positions, there was no lexical learning when listeners were exposed to ambiguous onset words (Jesse & McQueen, 2011). However, Cutler, McQueen, Butterfield, and Norris (2008) showed that legal word-onset phonotactic information can induce recalibration, presumably because this type of information can be used immediately whereas lexical knowledge about the word is not yet available when hearing the ambiguous onset. Moreover, lexical retuning is not restricted to a listener's native language as the English fricative theta ([θ] as in 'bath') presented in a Dutch f- or s-biasing context induced lexical learning (Sjerps & McQueen, 2010).

1.5.2 - Lipread induced- versus lexically-induced recalibration

So far, these data fit well with studies on lipread-induced recalibration but there was one remarkable difference: The duration of the reported aftereffects. Whereas lipread-induced recalibration was found to be fragile and short-lived (in none of tests did it survive more than 6-12 test trials, see van Linden & Vroomen, 2007; Vroomen & Baart, 2009b; Vroomen, et al., 2004), two studies on lexical-induced recalibration found that it was long-lived and resistant to change. Kraljic and Samuel (2005) demonstrated that recalibration of an ambiguous /s/ or /ʃ/ remained robust after a 25-minute delay. Moreover, it remained robust even after listeners heard canonical pronunciations of /s/ and /ʃ/ during the 25-minute delay and the only condition in which the effect became somewhat smaller, though not significantly, was when listeners heard canonical pronunciations of /s/ and /ʃ/ from the same speaker that they had originally adjusted to. In another study, Eisner and McQueen (2006) showed that lexical-induced recalibration remained stable over an even much longer delay (i.e., 12 hours) regardless of whether participants slept in the intervening time or not.

At this stage, one might conclude that, simply by their nature, lexical recalibration is robust and lipread recalibration is fragile. However, these studies were difficult to compare in a direct way because there were many procedural and item-specific differences. To examine this in more detail, van Linden and Vroomen (2007) conducted a series of experiments on lipread- and lexically-induced recalibration using the same procedure and test stimuli to check various possibilities. They used an ambiguous stop consonant halfway between /t/ or /p/ that could be disambiguated by either lipread or lexical information. For lipread recalibration, the auditory ambiguous

sound was embedded in Dutch non-words like ‘dikasoo?’ and dubbed onto the video of lipread ‘dikasoop’ or ‘dikasoot’; for lexical recalibration the ambiguous sound was embedded in Dutch p-words like ‘microscoo?’ (‘microscope’) or t- words like ‘idioo? (‘idiot’).

Across experiments, results showed that lipread and lexically recalibration effects were very much alike. The lipread aftereffect tended to be bigger than the lexical one, which was to be expected because lipreading has, in general, a stronger impact on sound processing than lexical information (Brancazio, 2004). Most importantly, though, both aftereffects dissipated equally fast, and there was thus no sign that lexical recalibration by itself was more robust than lipread-induced recalibration.

The same study also explored whether recalibration would become more stable if a contrast phoneme from the opposite category was included in the set of exposure items. Studies reporting long-lasting lexical aftereffects not only presented words with ambiguous sounds during exposure, but also presented filler words with non-ambiguous sounds taken from the opposite side of the phoneme continuum. For example, in the exposure phase of Norris et al. (2003) in which an ambiguous s/f sound was biased toward /f/, there were not only exposure stimuli like ‘witlo?’ that supposedly drive recalibration, but also contrast stimuli containing the nonambiguous sound /s/ (e.g., ‘naaldbos’). Such contrast stimuli might serve as an anchor or a comparison model for another stimulus and aftereffects thought to reflect recalibration might in this way be boosted because listeners set the criterion for the phoneme boundary in between the ambiguous token and the extreme one. The obtained aftereffect may then reflect the contribution of two distinct processes: One related to recalibration proper (i.e., a shift in the phoneme boundary meant to reduce the conflict between the sound and the context), the other to a strategic and long-lasting criterion setting operation that depends on the presence of an ambiguous phoneme and a contrast phoneme from the opposite category. Van Linden and Vroomen’s results (2007) showed that aftereffects indeed became substantially bigger if a contrast stimulus was included in the exposure set but crucially, aftereffects did not become more stable. Contrast stimuli thus boosted the effect, but did not explain why sometimes long-lasting aftereffects were obtained.

Another factor that was further explored was whether participants were biased in consecutive exposure phases towards only one, or both phoneme categories. One can imagine that if listeners are biased towards both a /t/ and /p/ (as was standard in lipread studies, but not the lexical ones), the boundary setting that listeners adopt may become fragile. However, this did not turn out to be critical: Whether participants were exposed to only one or both contexts, it did not change the size and stability of the aftereffect.

Of note is that lipread and lexical recalibration effect did not vanish when a 3-min silent interval separated the exposure phase from test. This finding indicates that recalibration as such is not fragile, but that other factors possibly related to test itself may explain why aftereffects dissipate quickly during testing. One such possibility might be that listeners adjust their response criterion in the course of testing and the two response alternatives are chosen about equally often. However, although this seems reasonable, it does not explain why, in the same test, selective speech adaptation effects remained stable in due course of testing (Vroomen, et al., 2004).

Yet another possibility is that recalibration needs time to consolidate and sleep might be a factor in this. Eisner and McQueen (2006) investigated this possibility and observed equal amounts of lexical-induced aftereffects after 12 hours, regardless of whether listeners had slept or not. Vroomen and Baart (2009b, see also Chapter 2) conducted a similar study on lipread-induced recalibration, including contrast phonemes to boost the aftereffect, and tested participants twice: immediately following the lipread exposure phase (as was standard) and after a 24-hour period during which participants had slept. The authors found large recalibration effects in the beginning of the test (the first 6 test trials), but they again quickly dissipated with prolonged testing (within 12 trials), and did not reappear after a 24-hour delay.

It may also be the case that the dissipation rate of recalibration depends on the acoustic nature of the stimuli. The studies that found quick dissipation used intervocalic and syllable-final stops that varied in place of articulation (/aba/-/ada/, and /p/-/t/), whereas others used fricatives (/f-s/ and /s-ʃ/; Eisner & McQueen, 2005; Kraljic, Brennan, & Samuel, 2008; Kraljic & Samuel, 2005) or syllable-initial voiced–voiceless stop consonants (/d-t/ and /b/-/p/; Kraljic & Samuel, 2006). If the stability of the phenomenon is depending on the acoustic nature of the cues (e.g., place cues might be more vulnerable), one may observe aftereffects to differ in this respect as well.

Another variable that may play a role is whether the same ambiguous sound is used during the exposure phase, or whether the token varies from trial-to-trial. Stevens (2007, Chapter 3) examined token variability in lexical recalibration using similar procedures as in Norris et al. (2003), but listeners were either exposed to the same or different versions of an ambiguous s/f sound embedded in s- and f-biasing words. His design also included contrast phonemes from the opposite phoneme category that should have boosted the effect. When the ambiguous token was constant, as in the original study by Norris et al. (2003), the learning effect was quite substantial on the first test trials, but it quickly dissipated with prolonged testing and in the last block (test trials 36-42), lexical recalibration had disappeared completely, akin to lipread-induced recalibration (van Linden & Vroomen, 2007; Vroomen, et al., 2004). When the sound

varied from trial-to-trial, the overall learning effect was much smaller and restricted to the f-bias condition, but the effect lasted longer.

Another aspect that may play a role is the use of filler items. Studies reporting short-lived aftereffects tended to use massed trials of adapters with either no filler items separating the critical items, or only a few contrast stimuli. Others, reporting long-lasting effects used lots of filler items separating the critical items (Eisner & McQueen, 2006; Kraljic & Samuel, 2005, 2006; Norris et al., 2003). Typically, about 20 critical items containing the ambiguous phoneme were interspersed among 180 fillers items. A classic learning principle is that massed-trials produce weaker learning effect than spaced trials (e.g., Hintzman, 1974). At present it remains to be explored whether recalibration is sensitive to this variable as well and whether it follows the same principle. One other factor that might prove to be valuable in the discussion regarding short- versus long-lasting effects is that extensive testing may override, or wash out, the learning effects (e.g., Stevens, 2007) because during the test, listeners might 're-learn' their initial phoneme boundary. Typically, in the Bertelson et al. (2003) paradigm, more test trials are used than in the Norris et al (2003) paradigm, possibly influencing the time course of the observed effects. For the time being, though, the critical difference between the short- and long-lasting recalibration effects remains elusive.

1.6 - Developmental aspects

Several developmental studies have suggested that integration of visual and auditory speech is already present early in life (e.g., Desjardins & Werker, 2004; Kuhl & Meltzoff, 1982; Rosenblum, Schmuckler, & Johnson, 1997). For example, four-month-old infants, exposed to two faces articulating vowels on a screen, look longer at the face that matches an auditory vowel played simultaneously (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999) and even 2-month old infants can detect the correspondence between auditory and visually presented speech (Patterson & Werker, 2003). However, it has also been demonstrated that the impact of lipreading on speech perception increases with age (Massaro, 1984; McGurk & MacDonald, 1976). Such a developmental trend in the impact of visual speech may suggest that lipreading is an ability that needs to mature, or alternatively that it requires linguistic experience, possibly because visible articulation is initially not well-specified. Exposure to audiovisual speech may then be necessary to develop phonetic representations more completely. Van Linden and Vroomen (2008) explored whether there is a developmental trend in the use of lipread information by testing children of two age groups, five-year-olds and eight-year-olds, on lipread-induced recalibration. Results showed that the older children learned to categorize the initially ambiguous speech

sound in accord with the previously seen lipread information, but this was not the case for the younger age group. Presumably, eight-year-olds adjusted their phoneme boundary to reduce the phonetic conflict in the audiovisual stimuli and this shift may occur in the older group but not the younger one because lipreading is not yet very effective at the age of five. Lexically-guided retuning however, has been observed in both six- and twelve-year-olds (McQueen, Tyler, & Cutler, in press).

Interestingly, poor lipreading skills may well be linked to poor reading skills (e.g., de Gelder & Vroomen, 1998; Mohammed, Campbell, Macsweeney, Barry, & Coleman, 2006; Ramirez & Mann, 2005) and because it is well-known that phonetic speech categories are less-well defined in developmental dyslexic- than fluent readers (e.g., Bogliotti, Serniclaes, Messaoud-Galusi, & Sprenger-Charolles, 2008; de Gelder & Vroomen, 1998; Godfrey, Syrdal-Lasky, Millay, & Knox, 1981; Vandermosten et al., 2010; Werker & Tees, 1987), Baart and Vroomen (Chapter 6) explored the possible relation between reading problems and lipread-induced recalibration by investigating recalibration aftereffects in 12 students with dyslexic reading problems. One could argue that an impaired ability to rely on the visual speech signal implies that some, and sometimes necessary, adjustments within the auditory system are not made, which might cause poorer defined auditory speech representations.

On the other hand however, auditory speech impairments could equally likely be the cause of the lipread problems as dyslexia-related auditory speech deficits are already present at birth. For example, newborns with familial risk for dyslexia display deviant brain activity when compared to non-risk infants when presented with synthetic /ba/, /da/ and /ga/ sounds (Guttorm, Leppanen, Tolvanen, & Lyytinen, 2003; Leppanen, Pihko, Eklund, & Lyytinen, 1999), which in turn is closely related to poorer receptive language skills and verbal memory in the following years of development (Guttorm et al., 2005).

Auditory identification of the /aba/-/ada/ continuum yielded shallower slopes for the dyslexic readers than for the fluent ones, indicating that /b/-/d/ categories were less well defined in the poor readers, in line with earlier reports (e.g., de Gelder & Vroomen, 1998; Godfrey et al., 1981; Werker & Tees, 1987). However, recalibration was alike for both groups indicating that the cross-modal learning mechanism that presumably underlies recalibration, is not affected by dyslexia.

Although lipreading is characterized by a developmental trend (Massaro, 1984), Teinonen, Aslin, Alku and Csibra (2008), were able to observe learning effects induced by lipread speech in young infants. They exposed 6-month-old infants to speech sounds from a /ba/-/da/ continuum. One group was exposed to audiovisual congruent mappings so that tokens from the /ba/-side of the continuum were combined

with lipread /ba/, and tokens from the /da/-side were combined with lipread /da/. Two other groups of infants were presented with the same sounds from the /ba/-/da/ continuum, but in one group all auditory tokens were paired with lipread /ba/, in the other group all auditory tokens were paired with lipread /da/. In the latter two groups, lipread information thus did not inform the infant how to divide the sounds from the continuum into two categories. A preference procedure revealed that infants in the former, but not in the two latter groups learned to discriminate the tokens from the /ba/-/da/ continuum. These results suggest that infants can use lipread information to adjust the phoneme boundary of an auditory speech continuum. Further testing though, is clearly needed so as to understand what critical experience is required and how it relates to lipread-induced recalibration in detail.

1.7 - Computational mechanisms

How might the retuning of phoneme categories be accomplished from a computational perspective? In principle, there are many solutions. All that is needed is that the system is able to use context to change the way an ambiguous phoneme is categorized. Recalibration may be initiated whenever there is discrepancy between the phonological representations induced by the auditory and lipread input, or for lexical recalibration, if there is a mismatch between the auditory input and the one expected from lexical information. Recalibration might be accomplished at the phonetic level by moving the position of the whole category, by adding the ambiguous sound as a new exemplar of the appropriate category, or by changing the category boundaries. For example, in models like TRACE (McClelland & Elman, 1986) or Merge (Norris, McQueen, & Cutler, 2000), speech perception is envisaged in layers where features activate phonemes that in their turn activate words. Here, one can implement recalibration as a change in the weights of the auditory feature-to-phoneme connections (Mirman, McClelland, & Holt, 2006; Norris et al., 2003).

Admittedly, though, the differences among these various possibilities are quite subtle. Yet, the extent to which recalibration generalizes to new exemplars might be of relevance to distinguish these alternatives. One observation is that repeated exposure to typical McGurk-stimuli containing a canonical sound, say non-ambiguous auditory /ba/ combined with lipread /ga/, does not invoke a retuning effect of the canonical /ba/ sound itself (Roberts & Summerfield, 1981). A ‘good’ auditory /ba/ thus remains a good example of its category, despite that there is lipread input repeatedly indicating that the phoneme belonged to another category. This may suggest that recalibration reflects a shift in the phoneme boundary, and thus only affecting sounds near that boundary,

rather than that the acoustic-to-phonetic connections are rewired on the fly, thus affecting all sounds, and in particular the trained ones.

In contrast with this view, though, there are also some data indicating the opposite. In particular, a closer inspection of the data from Shigeno (2002) shows that a *single* exposure to a McGurk-like stimulus AbVg, - here referred to as an 'anchor' - and followed by a target sound *did* change the quality of canonical target sound /b/ (see Figure 2 in Shigeno, 2002). This finding may be more in line with the idea of a 're-wiring' of feature-to-phoneme connections, or alternatively that this specific trained sound is incorporated in the new category. Clearly, though, more data are needed that specifically address these details.

There has also been a controversy about whether lexical recalibration actually occurs at the same processing level as immediate lexical bias. Norris et al. (2003) have argued quite strongly in favour of two types of lexical influence in speech perception: a lexical bias on phonemic decision-making that does not involve any form of feedback, and lexical feedback necessary for perceptual learning. Although there is a recent report supporting the idea of a dissociation between lexical involvement in on-line decisions and in lexical recalibration (McQueen et al., 2009) not all data support this distinction. That is, dissociating a bias (lipread or lexical) from recalibration has proven to be challenging: In fact, listeners who were strongly biased by the lipread or lexical context from the adapter stimuli (as measured in separate tests) also tended to show the largest recalibration effects (van Linden & Vroomen, 2007). Admittedly, this argument is only based on a correlation, and the correlation was at best marginally significant. Perhaps more relevant though, are the SWS findings (see Chapter 3) in which it was demonstrated that when lipread context did not induce a cross-modal bias - namely in the case where SWS stimuli were perceived as non-speech -, there was also no recalibration. Immediate bias and recalibration thus usually go hand-in-hand, and to claim that they are distinct, one would like to see empirical evidence in the form of an observed dissociation.

1.8 - Neural mechanisms

What are the neural mechanisms that underlie phonetic recalibration? The integration of auditory speech and a lipread- or lexical context has been extensively studied with brain imaging and electrophysiological methods (e.g., Callan et al., 2003; Calvert et al., 1997; Calvert & Campbell, 2003; Campbell, 2008; Colin et al., 2002; Holcomb & Neville, 1990; Klucharev, Möttönen, & Sams, 2003; Sams et al., 1991; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005). For instance, lipread speech context modulates auditory speech processing as early as 100

msec after stimulus onset as reflected by the attenuation and speeding-up of the N1 component in the ERPs (Besle, Fort, Delpuech, & Giard, 2004; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005) whereas lexically induced modulation of auditory speech processing is often reported to occur at around 400 msec (e.g., Holcomb & Neville, 1990). However, there is accumulative evidence that the early effects of lipread speech reflect low-level visual prediction (i.e., the anticipatory visual motion warns the listener about when a sound is going to occur) rather than higher-level phonetic integration that presumably occurs later in time (e.g., Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010). The experiment in Chapter 7 used SWS and demonstrated that the positive peak in the auditory ERP signal at 200 msec (i.e., the so-called P2) was attenuated by the lipread context only if the participants were aware of the speech-origin of the sounds (i.e., 'speech-mode') suggesting that the P2 potentially reflects phonetic binding.

However, in the case of phonetic recalibration, so far, only few studies have addressed the potential brain mechanisms involved in this process. Van Linden et al. (2007) used the mismatch negativity (MMN) as a tool to examine whether a recalibrated phoneme left traces in the evoked potential. The MMN is a component in the event-related potential (ERP) that signals an infrequent discriminable change in an acoustic or phonological feature of a repetitive sound (Näätänen, Gaillard, & Mäntysalo, 1978), and its latency and amplitude is correlated with the behavioural discriminability of the stimuli (Lang et al., 1990). The MMN is thought to be generated through automatic change detection and is elicited irrespective of sound-relevance for the participant's task (Näätänen, 1992; Näätänen, Paavilainen, Tiitinen, Jiang, & Alho, 1993). The MMN is not only sensitive to acoustic changes, but also to learned language specific auditory deviancy (Näätänen, 2001; Winkler et al., 1999). Van Linden et al. (2007) used a typical oddball paradigm to elicit a MMN so as to investigate whether lexical-induced recalibration penetrates mechanisms of perception at early pre-lexical levels, and thus affects the way a sound is heard. The standard stimulus (delivered in 82% of the trials) was an ambiguous sound halfway between /t/ and /p/ in either a t-biasing context 'vloo?' (derived from 'vloot', meaning 'fleet') or a p-biasing context 'hoo?' (derived from 'hoop', meaning 'hope'). For the deviant condition, the ambiguous sound was replaced by an acoustically clear /t/ in both conditions, so 'vloot' for the t-biasing context and 'hoot' (a pseudoword in Dutch) for the p-biasing context. If subjects had learned to 'hear' the sound as specified by the context, we predicted the perceptual change - as indexed by MMN - from /?/ → /t/ to be smaller in t-words than p-words, despite that the acoustic change was identical. As displayed in Figure 1b, the MMN in t-

words was indeed smaller than in p-words, thus confirming that recalibration might penetrate low-level auditory mechanisms.

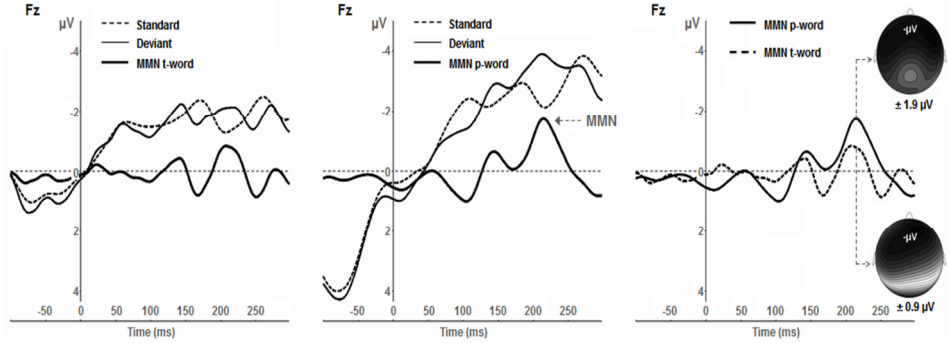

Figure 1b. *Grand-averaged waveforms of the standard, deviant and MMN at electrode Fz for the t-word condition (left panel) and p-word condition (middle panel) adapted from van Linden et al. (2007). The right panel shows the MMNs and their scalp topographies for both conditions. Voltage map ranges in μV are displayed below each map. The y-axis marks the onset of the acoustic deviation between /?/ and /t/.*

The second line of research concerned with potentially involved brain mechanisms used functional Magnetic Resonance Imaging (fMRI) to examine the brain mechanisms that drive phonetic recalibration (Kilian-Hütten, Valente, Vroomen, & Formisano, 2011; Kilian-Hütten, Vroomen, & Formisano, 2011). The original study by Bertelson et al. (2003) was adapted for the fMRI scanner environment. Listeners were presented with a short block of 8 audiovisual adapters containing the ambiguous /aba/-/ada/ sound dubbed onto the video of lipread /aba/ or /ada/ (A?Vb or A?Vd). Each exposure block was followed by 6 auditory test trials, consisting of event-related forced-choice /aba/-/ada/ judgments. Functional runs were analyzed using voxel-wise multiple linear regression (GLM) of the blood-oxygen level dependent (BOLD)-response time course. Brain regions involved in the processing of the audiovisual stimuli were identified by contrasting the activation blocks with a baseline. Moreover, a contrast based on behavioral performance was utilized to identify regions of interest (ROIs) whose activation during the recalibration phase would predict subsequent test performance (see also Formisano, De Martino, Bonte, & Goebel, 2008). Behaviorally, the results of Bertelson et al. (2003) were replicated in the fMRI environment, so more /aba/ responses after exposure to A?Vb than A?Vd. The auditory test stimuli elicited activation in regions within the anterior planum temporal adjacent to the posterior bank of Heschl's gyrus and sulcus, suggesting that pure perceptual interpretation of physically identical phonemes can be decoded from cortical activation patterns in early auditory areas (Kilian-Hütten, Valente, et al., 2011). As expected, the audiovisual

exposure blocks elicited activation in typical areas, including primary and extrastriate visual areas, early auditory areas, superior temporal gyrus and sulcus (STG/STS), middle and inferior frontal gyrus (MFG, IFG), pre-motor regions, and posterior parietal regions. Most interestingly, the BOLD-behaviour analysis identified a subset of this network (MFG, IFG, and inferior parietal cortex) whose activity during audiovisual exposure correlated with the proportion of correctly-recalibrated responses in the auditory test trials. Activation in areas MFG, IFG and inferior parietal cortex thus predicted the subjects' percepts of ambiguous sounds to be tested some 10 sec later (Kilian-Hütten, Vroomen, et al., 2011). Although these brain areas are also known to be involved in working memory (Jonides et al., 1998) phonetic recalibration does not seem to depend on working memory as such (Baart & Vroomen, 2010b, see also Chapter 5). The functional interpretation of these areas is to be explored further, but the activation changes may reflect trial-by-trial variations in participants' processing of the audiovisual stimuli, which in turn influence recalibration and later auditory perception.

1.9 - Summary and outline of this thesis

In this thesis, stability of lipread induced recalibration over time is investigated by measuring aftereffects immediately after exposure as well as 24 hours after exposure (**Chapter 2**). Although contrast stimuli were included, there was no indication that lipread induced recalibration lasted longer than 6 – 12 test-trials in the immediate test. In **Chapter 3**, recalibration is investigated with sine-wave speech in two groups of listeners; one in perceptual speech- and one in non-speech mode. Results indicated that recalibration only occurred in speech mode and was thus depending on whether the auditory and lipread signal were integrated into one phonetic event. Since spatial recalibration is bi-directional, as indicated by findings that a perceptual change in perceived sound location can be induced by a visually misaligned light as well as findings that a perceptual change in visual perception can be induced by a sound from a different location, the experiment in **Chapter 4** was set-up to investigate bi-directionality in audiovisual phonetic recalibration. The results indicated that perceived lipread identity was shifted towards the identity of the previously delivered auditory context, thus suggesting that phonetic audiovisual recalibration is indeed bi-directional. In **Chapter 5**, it was investigated whether phonetic recalibration is depending on working memory that possibly facilitates the longer-term perceptual adjustments but there was no indication that this is indeed the case. **Chapter 6** investigated phonetic recalibration in dyslexic readers because the auditory impairments in phonetic categorization, as are typical for dyslexic readers, might be linked to the ability to adjust the auditory speech system based on lipread context. Although auditory perception was

indeed impaired in the dyslexic group, recalibration was unaffected. In **Chapter 7**, it is suggested that audiovisual phonetic binding, presumably crucial for recalibration, takes place ~200 msec after auditory speech onset because ERP recordings obtained while participants were given sine-wave speech stimuli showed a lipread induced modulation of the P2 component, only when listeners were aware of the speech-like nature of the stimuli.

Chapter 2

Phonetic recalibration measured after 24 hours [2]

[2]Adapted from:

Vroomen, J. & Baart, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a twenty-four hour delay. *Language and Speech, 52*, 341-350.

2.1 - Abstract

Listeners hearing an ambiguous speech sound flexibly adjust their phonetic categories in accordance with lipread information indicating what the phoneme should be (recalibration). Here, we tested the stability of lipread-induced recalibration over time. Listeners were exposed to an ambiguous sound halfway between /t/ and /p/ that was dubbed onto a face articulating either /t/ or /p/. When tested immediately, listeners exposed to lipread /t/ were more likely to categorize the ambiguous sound as /t/ than listeners exposed to /p/. This aftereffect dissipated quickly with prolonged testing and did not reappear after a 24-hour delay. Audiovisual recalibration of phonetic categories is thus a fragile phenomenon.

2.2 - Introduction

Lipread information can help listeners by telling them how to interpret a speech sound that initially might be ambiguous. Imagine, for example, a speaker who pronounces an ambiguous sound intermediate between /b/ and /d/ in the context of the sentence "Could you please pass me the b/dutter." By looking at the speaker's face, listeners may notice that the lips were closed during pronunciation of the ambiguous sound, which is typical for /b/ but not for /d/. Moreover, there is also lexical knowledge informing the listener that the ambiguous sound should be /b/ rather than /d/, because "butter" but not "dutter" is a word in English. Numerous studies have shown that when listeners are asked to categorize the ambiguous sound, they do indeed use lipread and lexical information (Ganong, 1980; Sumby & Pollack, 1954). In addition, there is not only an immediate or on-line effect of the context on sound categorization, but there is also an aftereffect because the next time listeners hear the same sound, they have learned from the past and now perceive the initially ambiguous "b/d" sound as /b/ right away (Bertelson et al., 2003; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006, 2007; Norris et al., 2003; van Linden & Vroomen, 2007; Vroomen et al., 2007; Vroomen et al., 2004). The occurrence of this aftereffect demonstrates that listeners have adjusted the phonetic categories of their language so as to adapt to the new sound. What is at present unknown, though, is how long this adaptive shift lasts over time because some have reported that phonetic recalibration is fragile and dissipates within minutes (Stevens, 2007; van Linden & Vroomen, 2007; Vroomen et al., 2004), while others have found that recalibration is robust (Kraljic & Samuel, 2005) and can last for hours (Eisner & McQueen, 2006). Here, we further examined the robustness of phonetic recalibration by testing listeners immediately after exposure and then re-testing them after a 24-hour delay.

Lipread-induced recalibration was first demonstrated by Bertelson et al. (2003). They exposed listeners to an ambiguous sound intermediate between /aba/ and /ada/ (A?) dubbed onto a face articulating either /aba/ or /ada/ (A?Vb or A?Vd). Participants shortly exposed to A?Vb tokens reported in a subsequent auditory-only speech identification test more /aba/ responses than when exposed to A?Vd. This was taken as a demonstration that the visual information had shifted the interpretation of the ambiguous auditory phoneme in its direction. The same study also showed that when a non-ambiguous sound was dubbed onto a congruent face (AbVb or AdVd), the proportion of responses consistent with the visual stimulus decreased. Participants exposed to AbVb thus reported fewer /aba/ responses than when exposed to AdVd. This was interpreted as a sign of selective speech adaptation (Eimas & Corbit, 1973) in which it is the repeated presentation of a non-ambiguous speech sound 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 categorization trials. Selective speech adaptation probably reveals fatigue of some of the relevant processes, most likely acoustic or phonetic in nature, although criterion setting may also play some role (Samuel, 1986).

In a follow-up study, it was explored how long recalibration and selective speech adaptation would last over time (Vroomen et al., 2004). Participants were again exposed to audiovisual exposure stimuli that contained either the non-ambiguous or ambiguous auditory speech tokens (AbVb, AdVd, A?Vd, or A?Vb). Immediately after exposure, participants then categorized 60 auditory-only ambiguous speech tokens as /aba/ or /ada/. This allowed us to trace aftereffects as a function of time of testing. Results showed that aftereffects induced by ambiguous versus non-ambiguous sounds dissipated at different rates: Whereas recalibration effects were transient and lasted only about six-to-twelve test trials, selective speech adaptation effects were robust and were present even after 60 test trials. This difference in dissipation rates provided further evidence that the two phenomena resulted from distinct underlying mechanisms. It also showed that the transient nature of recalibration was not due to some particularity of the test itself (like participants trying to equally distribute the two response alternatives) because aftereffects of ambiguous and non-ambiguous sounds were tested in the same way.

In contrast with the transient nature of lipread-induced recalibration, studies on lexical recalibration have reported much more stable effects over time. Norris et al. (2003) were the first to demonstrate lexically-induced recalibration. They spliced an ambiguous fricative intermediate between /f/ and /s/ onto Dutch words normally ending in /s/ (e.g., radijs; radish) or /f/ (e.g., witlof; chicory). Exposure to the ambiguous sound embedded in words normally ending in /s/ (a /s/-biasing context) resulted in more /s/ responses on subsequent categorization trials if compared to the /f/-biasing context, thus revealing recalibration (or, in the authors' words, "perceptual learning"). When the ambiguous speech sound was spliced onto pseudo-words, no boundary shift was observed indicating that the shift was caused by lexical information proper. Others have since demonstrated the same phenomenon. For example, Kraljic and Samuel (2005) exposed listeners to a speaker whose pronunciation of the sound /s/ or /ʃ/ was ambiguous (halfway between /s/ and /ʃ/). Following an exposure phase, participants were tested for recalibration either immediately after exposure, or after a 25-min silent intervening task. Aftereffects were actually numerically bigger after the delay, indicating that simply allowing time to pass did not cause learning to fade. Even longer-lasting aftereffects were reported by Eisner and McQueen (2006). They exposed

listeners to a story in which they learned to interpret an ambiguous sound as /f/ or /s/. Results showed that perceptual adjustment measured after 12 hours was as robust as when measured immediately, and equivalent aftereffects were found when listeners heard speech from other talkers in the 12-hour interval or when they could sleep.

An obvious difference between studies that report robust versus fragile aftereffects is that the former used lexical information to induce recalibration, whereas fragile effects have been obtained with lipread speech. This difference in the way recalibration is induced, though, is unlikely to be relevant for the robustness of the effect because lipread effects tend, in general, to be bigger than lexical effects (e.g., Brancazio, 2004). This was confirmed in a study where lexical- and lipread-induced recalibration were compared directly with each other (van Linden & Vroomen, 2007). Listeners were exposed to an ambiguous sound halfway between /t/ and /p/ that was either dubbed onto a face articulating /t/ or /p/, or the sound was embedded in Dutch words normally ending in /t/ (e.g., 'groot', big) or /p/ (knoop, button). Following exposure to a lipread or lexical t- or p-bias, participants categorized auditory ambiguous tokens as /t/ or /p/. Results showed that the lipread-induced aftereffects tended to be somewhat bigger in size than the lexically-induced aftereffects, but both effects dissipated equally fast. It remains therefore unclear why some studies observed aftereffects to last for hours, while others reported fast dissipation.

One clue, though, may come from a procedural difference that has been demonstrated to play a role. Studies reporting robust aftereffects not only use the ambiguous sound that presumably drives recalibration (e.g., the ambiguous s/f-sound as embedded in the f-biasing context 'witlo/?/'; witlof = chicory), but listeners are also exposed to the non-ambiguous sound from the opposite phoneme category (in this example the non-ambiguous /s/ as embedded in radijs; radish). It has been demonstrated that the presence of this contrast phoneme during the exposure phase increases the size of the aftereffect (van Linden & Vroomen, 2007). There are various reasons, besides the already mentioned selective speech adaptation, why this may occur: the non-ambiguous contrast stimulus might, for example, serve as an anchor, or it might provide a comparison model for another stimulus. Aftereffects thought to reflect recalibration could in this way be boosted because listeners set the criterion for the phoneme boundary in between the ambiguous token and the extreme one. Alternatively, participants may also adopt a tendency to judge anything that is not a clear /s/ as an /f/. Either way, the obtained aftereffect will then reflect the contribution of two distinct processes. One is related to recalibration proper (i.e., a shift in the phoneme boundary that is meant to reduce the conflict between the auditory and lexical information), while

the other might be a strategic and longer lasting criterion setting operation that depends on the presence of two phonemes from opposing categories.

To explore this possibility, we addressed whether lipread-induced aftereffects become robust when contrast stimuli are presented during the exposure phase. Recalibration was induced by exposing participants to an ambiguous speech sound /?/ halfway between /t/ and /p/ that was combined with the non-ambiguous visual articulation of /t/ or /p/, (A?Vt for the t-biased group, and A?Vp for the p-biased group). In addition, non-ambiguous contrast stimuli from the opposing category were presented: ApVp for the t-biased group, and AtVt for the p-biased group. Following exposure to these stimuli, participants categorized ambiguous speech sounds from a /t/-/p/ continuum immediately after exposure and – to examine whether the effects were robust – were re-tested after a 24-hour delay.

2.3 - Method

2.3.1 - Participants

Twenty native speakers of Dutch (18-25 years old) with normal hearing and normal seeing participated. Half of them was biased towards /t/, the other towards /p/.

2.3.2 - Materials

The same stimuli were used as in van Linden and Vroomen (2007). Stimulus creation started with a video and audio recording of a male native speaker of Dutch. An auditory ambiguous sound intermediate between /t/ and /p/, henceforth /?/, was created using the Praat speech editor (http://www.praat.org). The /?/ was created from a recording of /ot/ of which the second (F2) and third (F3) formant were varied so as to create a 10-step /ot/–/op/ continuum. The steady state-value of the F2 in the vowel was 950 Hz and 72 ms in duration. The transition of the F2 was 45 ms, and its offset frequency varied from 1123 Hz for the /t/-endpoint to 600 Hz for the /p/-endpoint in ten equal Mel steps. The F3 had a steady state value of 2400 Hz in the vowel, and the offset frequency of the transition varied from 2350 Hz for the /t/- endpoint to 2100 Hz for the /p/-end point in ten equal Mel steps. The silence before the final release of the stop consonant was increased in 6 ms steps from 22 ms for the /t/-endpoint to 82 ms for the /p/-endpoint. The waveforms of the aspiration part of the final release of /p/ and /t/ (134 ms) were mixed from natural /p/ and /t/ bursts in relative proportions to each other. The resulting continuum sounded natural with no audible clicks.

For the exposure stimuli, the ambiguous sound /?/ was spliced into recordings of four different pseudowords such as woo/?/ (‘woot’ and ‘woop’ are both non-words).

The audio was then dubbed on the video of the face that articulated either 'woop' or 'woot'. For the auditory test tokens, /?/ was spliced into the pseudoword soo/?/.

2.3.3 - Procedure

Participants were tested individually in a soundproof booth. Half of the participants was biased towards /t/ and exposed to A?Vt and ApVp; the other half of the participants was biased towards /p/ and 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 speakers (JBL Media 100WH/230) placed on the left and right of the monitor. The video fragments were 10 x 9.5 cm in size, and were shown against a black background. Loudness peaked at 70 dBa. A regular keyboard was used for data-acquisition. During the test, participants were instructed to press the p-key upon hearing 'soop' and t-key upon hearing 'soot'.

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.

2.3.3.1 - Calibration

For each individual participant, it was determined which token was the most ambiguous one of the continuum. All test tokens were presented ten times in pseudo random order and participants were asked to indicate whether they heard 'soot' or 'soop'. The 50% crossover point was then determined via a logistic procedure. The token nearest this point served as the most ambiguous stimulus /?/ for subsequent testing.

2.3.3.2 - Training

To acquaint participants with the test procedure, they categorized the most 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 20 times in pseudo-random order.

2.3.3.2 - Exposure - test

Participants were presented five blocks of 16 exposure stimuli followed by 60 test trials. In the t-biased condition, each exposure block contained eight A?Vt stimuli and eight contrast stimuli ApVp in random order. In the p-biased condition, each exposure block contained eight A?Vp and eight AtVt stimuli. To ensure that participants were watching the monitor during the exposure phase, catch trials were included consisting of the 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 of the five exposure blocks was immediately followed by 60 auditory-only test trials. In the test phase, the three test tokens (soo/?−1/, soo/?/, and soo/?+1/) were presented 20 times in counterbalanced order. Participants were asked to indicate whether they heard 'soop' or 'soot' by pressing the 'p' or 't' key.

2.3.3.3 - Re-test after a 24-hour delay

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

2.4 - 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-biased group and 51% in the p-biased group, indicating that the proportion of /t/- and /p/-responses before exposure was alike in both groups. During exposure, 99% of the catch trials were detected, indicating that participants kept their eyes fixed on the monitor.

To measure aftereffects and their dissipation, the test trials were binned into 10 serial positions. Each position represented the mean average number of /t/-responses on a total of 30 test-trials (6 consecutive test-trials in each of the five test-blocks) The group-averaged proportion of /t/-responses is shown in Figure 2a (immediate test) and Figure 2b (delayed test). High values indicate more /t/- and thus less /p/-responses.

Aftereffects were calculated as in previous studies by subtracting the proportion of /t/-responses following /p/-bias from /t/-bias. Figure 2c shows the group-averaged difference for the immediate- and delayed test.

For the immediate test, a 2 (/t/- vs. /p/-bias) x 10 (test token position) ANOVA on the proportion of /t/-responses showed a significant main effect of exposure condition, $F(1,18) = 12.14$, $p < .003$, because there were, as expected, more /t/-responses following /t/-bias than /p/-bias. This is the basic recalibration effect. The interaction with test token position was also significant, $F(9,162) = 7.59$, $p < .001$, as the difference between the two groups dissipated with prolonged testing. Separate t-tests (Bonferroni corrected for multiple comparisons) showed that aftereffects were bigger than zero up to test token position 2, thus representing the first 12 test trials. On test trials 1 - 6, the /t/-biased group gave a substantial 48% more /t/-responses than the /p/-biased group, and on test trials 7 - 12 this difference was still 30%.

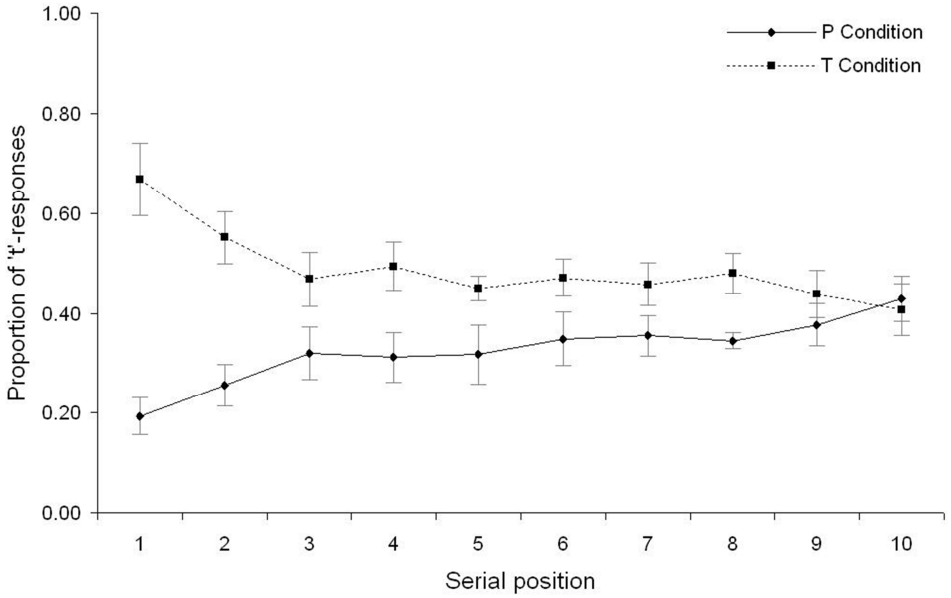

Figure 2a. *Proportion of /t/-responses as a function of the serial position in the immediate test. Error bars represent 1 standard error of the mean.*

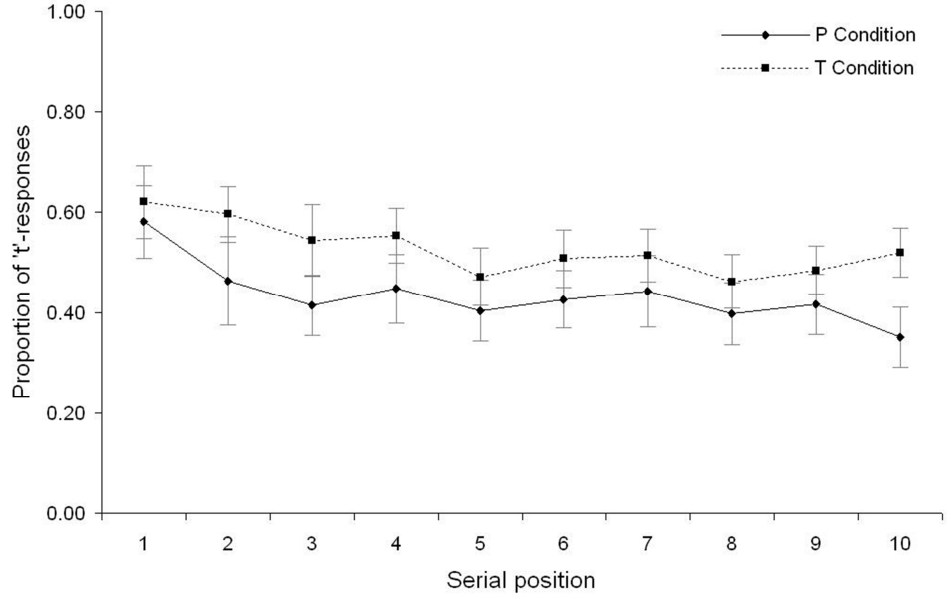

Figure 2b. *Proportion of /t/-responses as a function of the serial position in the delayed test. Error bars represent 1 standard error of the mean.*

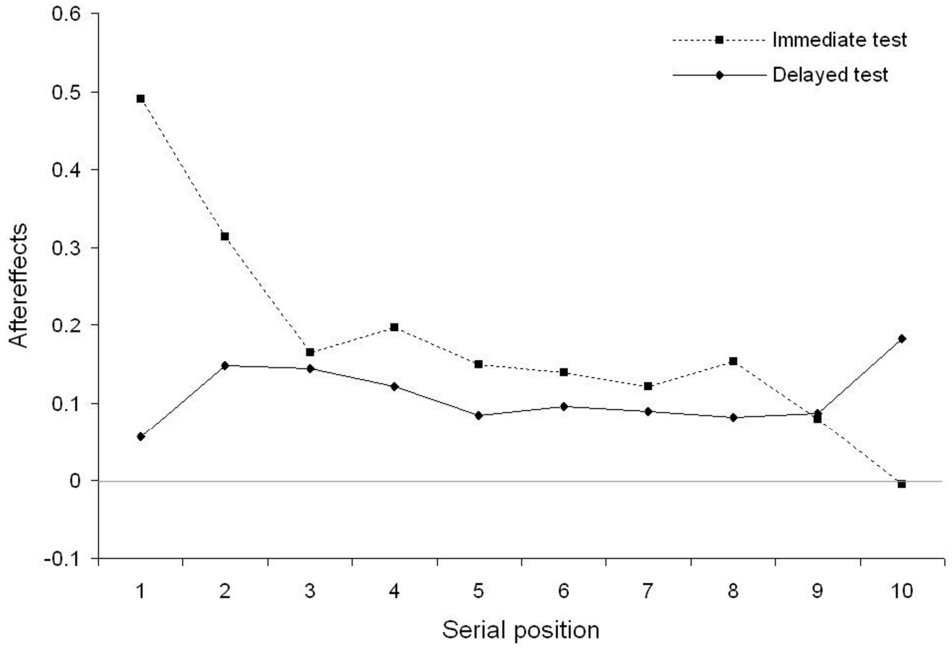

Figure 2c. *Aftereffects as a function of the serial position in the test.*

The same ANOVA on the data of the delayed test showed that after 24 hours, none of those effects was significant anymore. There was no overall difference between the /t/- and /p/-biased groups, $F(1,18) = 1.72$, $p < .206$, and the interaction with test token position was also not significant, $F < 1$. Thus, despite substantial aftereffects in the immediate test, they dissipated fast and did not survive the delay.

Finally, we examined aftereffects per test block, because in delayed testing there might have been an aftereffect in the first test block only. In the 5 (block) x 2 (/t/- vs. /p/-bias) x 10 (test token position) ANOVA, there was no main effect of block in the immediate and delayed test, $F(4,72) = 1.08$, $p < .372$ and $F(4,72) < 1$, respectively, and block did not interact with any of the other factors, all p's > .15. Visual inspection confirmed that aftereffects were essentially the same across all five test blocks.

2.5 - Discussion

Participants were biased to categorize an ambiguous speech sound as /t/ or /p/ by using two different kinds of exposure stimuli. On the one hand, we exposed participants to an auditory ambiguous sound combined with non-ambiguous lipread speech (A?Vt for the t-biased group and A?Vp for the p-biased group). Presumably, for these stimuli it is the lipread information that informs listeners on how to interpret the ambiguous sound. The phonetic conflict between the heard and lipread information thus

induces a shift in the phoneme boundary that reduces the conflict (i.e., recalibration proper). This shift can in subsequent testing be observed as an aftereffect. Secondly, participants were also exposed to contrast stimuli containing a non-ambiguous sound from the opposing phoneme category (ApVp for the t-biased group and AtVt for the p-biased group). The presence of this contrast phoneme was expected to boost and possibly stabilize aftereffects as it might help in settling where the phoneme boundary should be. Exposure to both kinds of stimuli indeed resulted in a substantial aftereffect (i.e., a 46% difference), but only on the first test token positions. With prolonged testing, the aftereffect dissipated quickly and it did not reappear after a 24-hour delay. Apparently, the presence of contrast stimuli did thus not stabilize aftereffects.

This result raises the question what it is that drives phonetic representations to be re-adjusted back to normal that quickly. This is a particularly intriguing question if it is realized that others have observed (lexical) recalibration to be extremely robust against various unlearning conditions whereby listeners even heard 'good' examples of previously adjusted tokens (Kraljic & Samuel, 2005). The simplest potential mechanism might be time itself whereby perceptual adjustments just 'fade away'. In the absence of any speech input, phonetic representations would then revert to their prior settings.
Previously, though, we showed, with the same stimuli as used here, that this does not obtain for lipread-induced recalibration because aftereffects did not become smaller when a three-minute silent interval intervened between the exposure phase and the test (van Linden & Vroomen, 2007, Experiment 4). Lipread-induced recalibration is thus not fragile as such.

Another possibility is that the test procedure itself induces dissipation. The test involves a large number of trials in which ambiguous sounds are presented. It might be that listeners adjust their response criterion in the course of testing such that the two response alternatives are chosen about equally often. One would then expect the test to be most sensitive on the first few trials, while differences between conditions will become washed out with prolonged testing. We and others have indeed observed that aftereffects become smaller with prolonged testing (see, e.g., Kraljic & Samuel, 2006). However, against this interpretation of response equilibration is the finding that in a previous study, there was no dissipation of aftereffects caused by selective speech adaptation, despite the fact that the same test as here was being used (Vroomen et al., 2004). Thus, after being exposed to, say, the non-ambiguous tokens AbVb, participants were less likely to report /b/ during the whole test. The test itself does thus not induce response equilibration. One notable difference, though, between selective speech adaptation and recalibration is that in the exposure phase of selective speech adaptation, no ambiguous speech sounds are presented whose phoneme boundary is shifted towards

one or the other side of the continuum. For recalibration, it might thus be that the shift in the phoneme boundary by itself causes participants to be flexible during the test as well. Recalibration effects would, on this view, thus be fragile because participants are in a 'shifting-mood'. Further tests, though, are needed to examine this speculation more thoroughly and explore the conditions under which recalibration remains stable or returns to normal again.

Another potentially important factor affecting the stability of recalibration may be the acoustic or phonetic nature of the stimulus that is adapted. Previous studies either used fricatives (/f/–/s/, /s/–/ʃ/) or stop consonants (/p/-/t/, /b/-/d/, and /d/-/t/). Fricatives tend to produce large shifts that are long-lasting, and are primarily speaker- or token-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007). Stop consonants, though, tend to produce smaller shifts that do not seem to last as long (Kraljic & Samuel, 2006; van Linden & Vroomen, 2007), but that generalize across speakers (Kraljic & Samuel, 2006, 2007). Further tests are required to explore whether the acoustic nature of the stimuli explains the difference in stability over time.

Another aspect that may play a role is the use of filler items. One of the classic learning principles is that massed trials produce a weaker learning effect than spaced trials (Ebbinghaus, 1885). In our previous studies, we always presented massed trials of adapters with either no filler items separating the critical items, or - as in the present study - only a few contrast stimuli (Bertelson et al., 2003; van Linden, et al. 2007; van Linden & Vroomen, 2007, 2008; Vroomen et al., 2007; Vroomen et al., 2004). Others, though, reporting long-lasting effects used lots of filler items separating each of the critical items (Eisner & McQueen, 2006; Kraljic & Samuel, 2005, 2006; Norris et al., 2003). Typically, about 20 critical items containing the ambiguous phoneme were interspersed among 180 fillers items. At present, it remains to be explored whether recalibration of phonetic categories is sensitive to this variable and whether it follows the same classic learning principle.

To conclude, we found that aftereffects induced by lipread recalibration were fragile despite the presence of contrast phonemes during the exposure phase. The size of the aftereffect was most boosted by the presence of contrast stimuli, but these stimuli did not make the effect more robust. The robustness of lexical aftereffects reported by others (Eisner & McQueen, 2006; Kraljic & Samuel, 2005) are, most likely, therefore not caused by the presence of contrast stimuli as such.

Chapter 3

Phonetic recalibration with artificial speech sounds [3]

[3]Adapted from:

Vroomen, J. & Baart, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition, 110*, 254 - 259.

3.1 - Abstract

Upon hearing an ambiguous speech sound dubbed onto lipread speech, listeners adjust their phonetic categories in accordance with the lipread information (recalibration) that tells what the phoneme should be. Here we used sine wave speech (SWS) to show that this tuning effect occurs if the SWS sounds are perceived as speech, but not if the sounds are perceived as non-speech. In contrast, selective speech adaptation occurred irrespective of whether listeners were in speech or non-speech mode. These results provide new evidence for the distinction between a speech and non-speech processing mode and they demonstrate that different mechanisms underlie recalibration and selective speech adaptation.

3.2 - Introduction

A critical question about speech is whether specialized processors are responsible for the coding of the acoustic signal in phonetic segments (Liberman & Mattingly, 1985) or whether speech is perceived as all other sounds (Massaro, 1987). A clear demonstration of the existence of a speech versus non-speech mode was provided by Remez et al. (1981) using sine-wave speech (SWS). In SWS, the natural richness of the auditory signal is reduced to a few sinusoids (usually three) that follow the centre frequency and the amplitude of the first three formants. These stimuli sound highly artificial, and most naïve subjects perceive them as 'non-speech' sounds like whistles or sounds from a science fiction movie. Typically, though, once subjects are told that these sounds are actually derived from speech, they cannot switch back to a non-speech mode again and continue to hear the sounds as speech. Functional brain imaging studies have provided converging evidence that for listeners in speech mode, there is stronger activity in the left superior temporal sulcus than for listeners in non-speech mode (Möttönen et al., 2006). Moreover, if SWS sounds are combined with lipread speech, naïve subjects in non-speech mode show no or only negligible intersensory integration (lipread information biasing speech sound identification), while subjects who learned to perceive the same auditory stimuli as speech do integrate the auditory and visual stimuli in a similar manner as natural speech (Tuomainen et al., 2005).

Previous studies demonstrating the speech/non-speech mode distinction had to rely on the immediate subjective report that the SWS stimuli were actually perceived as speech or non-speech. Here, we demonstrate that there are also indirect effects using two distinct phenomena that we hypothesized to be differently sensitive as to whether perceivers were in speech or non-speech mode, namely recalibration of phonetic categories and selective speech adaptation. Recalibration of phonetic categories is a tuning effect that occurs when a phonetically ambiguous speech sound is combined with lipread speech. While being exposed to such an audiovisual stimulus, participants adjust the phoneme boundary and learn to categorize the initially ambiguous speech sound in accordance with the simultaneously presented lipread speech. This can be demonstrated in a subsequent auditory-only test where listeners identify the ambiguous sound. For example, if an ambiguous sound halfway between /b/ and /d/ is dubbed onto lipread /b/, then participants are more likely to categorize the ambiguous sound as /b/. Presumably, recalibration is induced by the deviance between the heard and lipread information that the brain tries to minimize by shifting the phoneme boundary (Bertelson et al., 2003; van Linden & Vroomen, 2007; van Linden & Vroomen, 2008; Vroomen et al., 2007; Vroomen et al. 2004).

Selective speech adaptation, first demonstrated by Eimas and Corbit (1973), is different from recalibration in that it does not depend on a conflict between two information sources, but rather depends on the repeated presentation of a particular speech sound by itself that causes a reduction in the frequency with which that token is reported in subsequent identification trials. Since its introduction, many questions have been raised about the nature underlying this effect. Originally, it was thought to reflect a fatigue of some hypothetical 'linguistic feature detectors', but others argued that it reflects a shift in criterion (Diehl et al., 1978), or a combination of both (Samuel, 1986). Others however (e.g., Sawusch, 1977), showed that the size of selective speech adaptation depends upon the degree of spectral overlap between the adapter and test sound, and that most, if not all of the effect is auditory rather than phonetic. A similar conclusion was reached by Roberts and Summerfield (1981). They exposed listeners to audiovisual congruent (auditory /b/ with lipread /b/) or incongruent adapter stimuli (auditory /b/ with lipread /g/) and obtained similar aftereffects, despite the fact that the adapters were perceived differently. Selective adaptation thus mainly depends on the acoustic nature of the adapter, and not the lipread component or the phonetic percept (see also Saldaña & Rosenblum, 1994).

Here, we examined whether recalibration and selective speech adaptation occurs with SWS stimuli, and whether the effects would differ for listeners in speech versus non-speech mode. We hypothesized that lipread-induced recalibration occurs if, and only if, perceivers are in speech mode but not in non-speech mode because in non-speech mode there is no intersensory integration (Tuomainen et al., 2005) and hence no phonetic conflict between sight and sound that would induce recalibration. We thus assumed that recalibration occurs to the extent that conflicting information sources are referring to the same event. If listeners are in speech mode, heard and lipread inputs are combined into a single phonetic representation, but not so if listeners are not under the impression that the auditory and visual signals refer to separate events. Selective adaptation, though, may occur for listeners in speech and non-speech mode, assuming that this phenomenon depends on some low-level acoustic factor and not the phonetic interpretation of the sound (Roberts & Summerfield, 1981).

To test these hypotheses, we created an SWS continuum between /omso/ and /onso/. Participants were trained to categorize the two auditory endpoints of this continuum as /omso/ or /onso/ for the speech group, or as '1' or '2' for the non-speech group. Once participants reliably discriminated the two sounds, they were exposed to audiovisual adapter stimuli intended to induce recalibration or selective speech adaptation and then tested. To induce recalibration, we used audiovisual adapters containing the most ambiguous SWS token of the continuum halfway between /omso/

and /onso/ (henceforth /A?/ for 'Auditory ambiguous') dubbed onto a video recording of the speaker articulating /omso/ or /onso/ (A?Vomso and A?Vonso). Following a short exposure phase, auditory-only test trials were given in which participants identified the SWS tokens from the middle of the continuum. For participants in speech mode, we expected the ambiguous tokens to be labeled in accordance with the previously seen lipread adapter, so more /onso/-responses after exposure to A?Vonso than A?Vomso. No such difference was expected for the non-speech group, because lipread speech should not affect the auditory tokens if they are labeled as non-speech (Tuomainen et al., 2005).

To induce selective adaptation, we used audiovisual adapters containing the endpoint tokens of the /omso/-/onso/ continuum, and dubbed these onto congruent video recordings of the speaker. Participants were thus exposed to AomsoVomso and AonsoVonso. Due to the non-ambiguous acoustic nature of the sound, we expected to observe contrastive aftereffects irrespective of whether participants were in speech or non-speech mode, so more /onso/- or '2'-responses after exposure to AomsoVomso and more /omso/ or '2'-responses after AonsoVonso.

3.3 - Method

3.3.1 - Participants

Twenty-four native speakers of Dutch (first-year students) participated. Half of them were trained in speech mode, the other half in non-speech mode.

3.3.2 - Stimuli

Stimulus creation started from the original recording of natural /omso/ and /onso/ tokens previously used by Tuomainen et al. (2005). Using the Praat-programme (Boersma & Weenink, 2005), a seven-point continuum between /omso/ and /onso/ was created by changing the second (F2) and third (F3) formants in equal steps. The steady state value of the F2 in the initial vowel was 780 Hz and lasted 140 ms for both endpoints. The transition of the F2 in the nasal was 50 ms, and its offset frequency varied from 1800 Hz for the /onso/-endpoint to 680 Hz for the /omso/-endpoint in equal Mel steps. The F3 had a steady state value of 2500 Hz in the vowel, and the offset frequency of the transition varied from 2500 Hz for the /onso/-endpoint to 2250 Hz for the /omso/-end point. This resulted in a natural sounding seven-point /omso/-/onso/ continuum. Pilot tests showed (N = 16) that the middle (fourth) stimulus was also the most ambiguous one (see Figure 3a).

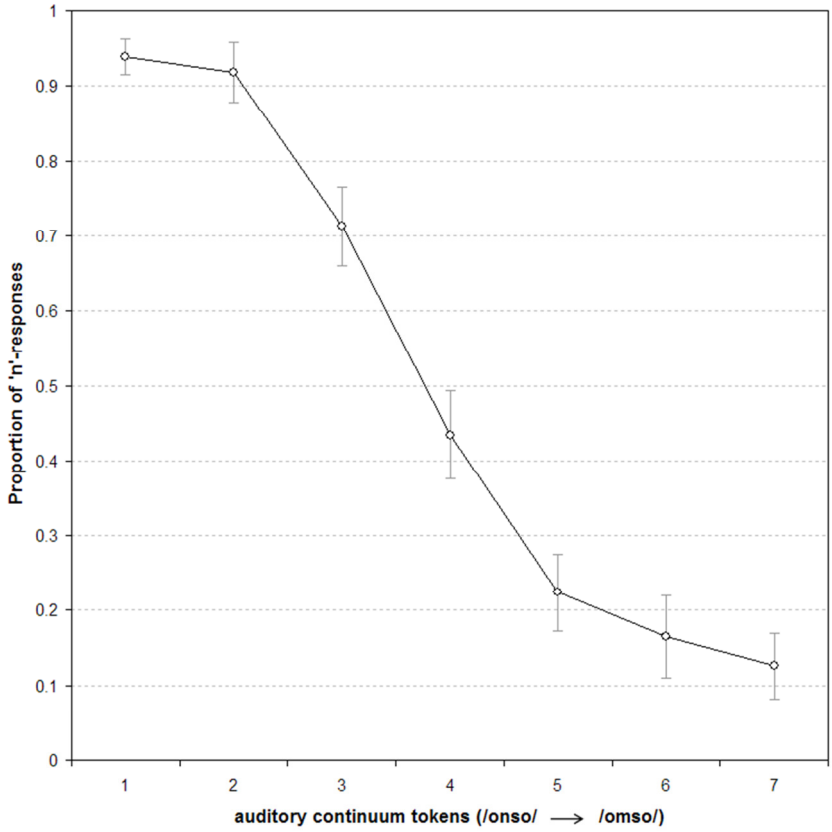

Figure 3a. *Mean proportion of /onso/ responses of the original synthetic continuum. Error bars represent one standard error of the mean.*

The tokens of the thus created continuum were transformed into SWS sounds using a script from C. Darwin available on the internet (http://www.biols.susx.ac.uk/home/Chris_Darwin /Praatscripts /SWS). Three-tone SWS stimuli were created with time varying sine waves for the three lowest formants (Figure 3b). These SWS stimuli where then dubbed onto the video recording of the speaker (29.97 frames per s., 22 x 15 cm) articulating either /omso/ or /onso/, preserving the natural timing between the audio and video. This resulted in four audiovisual adapter stimuli: A?Vomso and A?Vonso (to induce recalibration) and AomsoVomso and AonsoVonso (to induce selective speech adaptation). The sound level of the stimuli peaked at 79 dBa when measured at ear level.

To ensure that participants were looking at the screen during adaptation, participants had to detect a small white dot that appeared for 100 ms on the upper lip of

the speaker. Participants had to press a special key upon appearance of such an occasional catch trial.

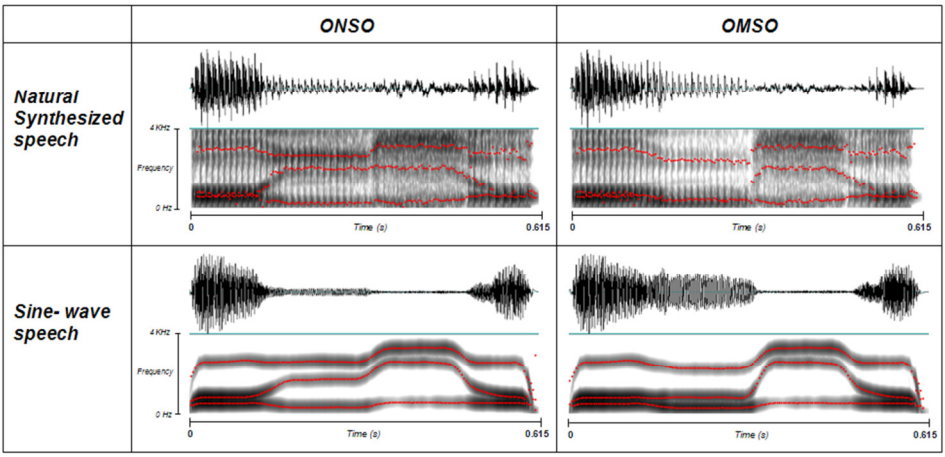

Figure 3b. *Waveforms and corresponding spectograms of the endpoints of the synthesized continuum and their sine wave replicas. Formants (F1, F2 and F3) are represented by dotted lines.*

3.3.3 - Procedure and design

Participants were tested individually in a sound attenuated and dimly lit room at 70 cm distance from a 17-inch CRT monitor. They were first acquainted with the SWS tokens and learned to categorize the endpoints as /omso/ and /onso/ for the speech group, or as '1' and '2' for the non-speech group. The two endpoints were delivered 48 times in pseudorandom order with immediate feedback. Participants continued training without corrective feedback until a learning criterion was met (12 consecutively correct answers). Two participants (one in speech mode, the other in non-speech mode) failed to meet this criterion after a predetermined time limit and were replaced. The learning criterion was reached after 33.0 trials for the speech group, and 24.3 trials for the non-speech group; $t(22) = .86$, $p = .40$. From the start, both groups were thus equally good in discriminating the auditory SWS endpoints.

3.3.4 - Adapter-test blocks

Similar procedures were used as in Bertelson et al. (2003, Experiment 2). Participants were repeatedly exposed to short blocks of audiovisual adapter stimuli immediately followed by auditory-only test trials. Each adapter block contained eight consecutively presented adapter stimuli (either A?Vomso, A?Vonso, AomsoVomso, or AonsoVonso, ISI = 425 ms) followed by six test trials. In the test, the most ambiguous

SWS token (A?) and the more /onso/-like (A?-1) and /omso/-like stimulus (A?+1) of the continuum were presented twice. Participants pressed a designated key upon perceiving /omso/ (or '1') or /onso/ (or '2'). Participants were exposed to eight blocks of each adapter (32 adapter-test blocks in total), all presented in pseudorandom order.

3.3.5 - Goodness ratings of the adapters

In the final part, participants rated the auditory quality of the audiovisual adapter stimuli. Each adapter was presented six times in pseudorandom order and participants rated the goodness of the sound on a seven-point Likert scale with '1' for a clear /omso/ or '1', and '7' for a clear /onso/ or '2'. Finally, participants in the non-speech group were asked whether they had noticed that the SWS stimuli originated from actual speech. Three reported to have heard spoken syllables (though not /omso/ and /onso/) and were replaced by others.

3.4 - Results

3.4.1 - Catch trials

Performance on catch trials was almost flawless (99% correct for the speech group and 96% correct for the non-speech group) indicating that participants were indeed looking at the video during exposure to lipread speech.

3.4.2 - Goodness ratings of adapters

The goodness ratings of the audiovisual adapters were analyzed first to ensure that they were perceived as intended. As in Tuomainen et al. (2005), lipreading had a strong impact on the ambiguous SWS sound if the sound was perceived as speech, but not if perceived as non-speech (see Table 3.1). In the 2 (speech/non-speech mode) x 2 (ambiguity of adapter sound) x 2 (lipread /omso/ or /onso/) overall ANOVA, the critical interaction between mode, ambiguity of adapter sound, and lipread adapter was highly significant, $F(1,22) = 16.34$, $p < .002$ ($\eta^2 = .43$). A separate ANOVA for adapter stimuli with ambiguous sounds (A?Vomso, A?Vonso) showed the main effect of lipreading, $F(1,22) = 19.77$, $p < .001$ ($\eta^2 = .47$), interacted with speech mode, $F(1,22) = 19.80$, $p < .001$ ($\eta^2 = .47$). Separate t-test confirmed that lipreading affected the quality of the ambiguous sound if listeners were in speech mode (a 2.79 bias, $t(11) = 4.98$, $p < 0.001$), but not so if listeners were in non-speech mode (0.00 bias, testing unneeded). The ANOVA for adapter stimuli with non-ambiguous sounds (AomsoVomso or AonsoVonso) showed there was a significant stimulus effect, $F(1,22) = 487.68$, $p < .001$ ($\eta^2 = .96$), but no interaction with speech mode ($F < 1$). Non-ambiguous sounds were thus equally distinct for both groups.

Mode	Exposure sound	Lipread adapter		Lipread bias
		/onso/	/omso/	
Speech Mode	Ambiguous	5.27	2.48	2.79
	Non-ambiguous	6.50	1.69	4.81
Non-speech Mode	Ambiguous	4.19	4.19	0.00
	Non-ambiguous	6.37	1.74	4.63

Table 3.1. *Goodness ratings of the audiovisual adapters. Lipread bias was calculated by subtracting /omso/ ratings from /onso/.*

3.4.3 - Test trials

Performance on test trials following exposure to the different adapters is presented in Figure 3c. We also computed aftereffects as in previous studies (Bertelson et al., 2003) by subtracting the proportion of /onso/-responses following exposure to /omso/ from /onso/ pooling over the three test tokens (see Table 3.2). As is clearly visible, for the speech group there was recalibration and selective speech adaptation, while for the non-speech group there was only selective adaptation with no sign of recalibration. In the 2 (speech/non-speech mode) x 2 (ambiguity of adapter sound) x 2 (lipread /omso/ or /onso/) x 3 (Auditory test token) overall ANOVA, there was a main effect of ambiguity of the adapter sound $F(1,22) = 5.80$, $p < .025$ ($\eta^2 = .21$), because there were more /onso/-responses after exposure to non-ambiguous adapter sounds, and a main effect of auditory test token, $F(2,44) = 63.62$, $p < .001$ ($\eta^2 = .74$), because there were more /onso/- (or '2'-) responses for sounds from the /onso/- than /omso/-side of the continuum. The interaction between the ambiguity of the adapter sound and lipread speech was significant, $F(1,22) = 31.57$, $p < .001$ ($\eta^2 = .59$), because there were more /onso/-responses after exposure to A?Vonso than A?Vomso, (i.e., recalibration), but less /onso/-responses following exposure to AonsoVonso than AomsoVomso (i.e., selective speech adaptation). Most important, the size of this effect differed for the speech and non-speech group as reflected in a significant second-order interaction, $F(1,22) = 5.14$, $p < .034$ ($\eta^2 = .19$).

Separate t-tests confirmed that for the speech group, there were 14% more /onso/-responses after exposure to A?Vonso than A?Vomso, $t(11) = 3.96$, $p < .002$ (one-sided, as there was a clear prediction), while there were 19% fewer /onso/-responses

after exposure to AonsoVonso than AomsoVomso, t(11) = 2.40, $p < .036$. For the non-speech group, there was no difference (0%) between A?Vonso and A?Vomso,; t(11) = .17, $p < .87$, whereas there were 13% fewer /onso/-responses after exposure to AonsoVonso than AomsoVomso, t(11) = 4.59, $p < .001$ (selective speech adaptation).

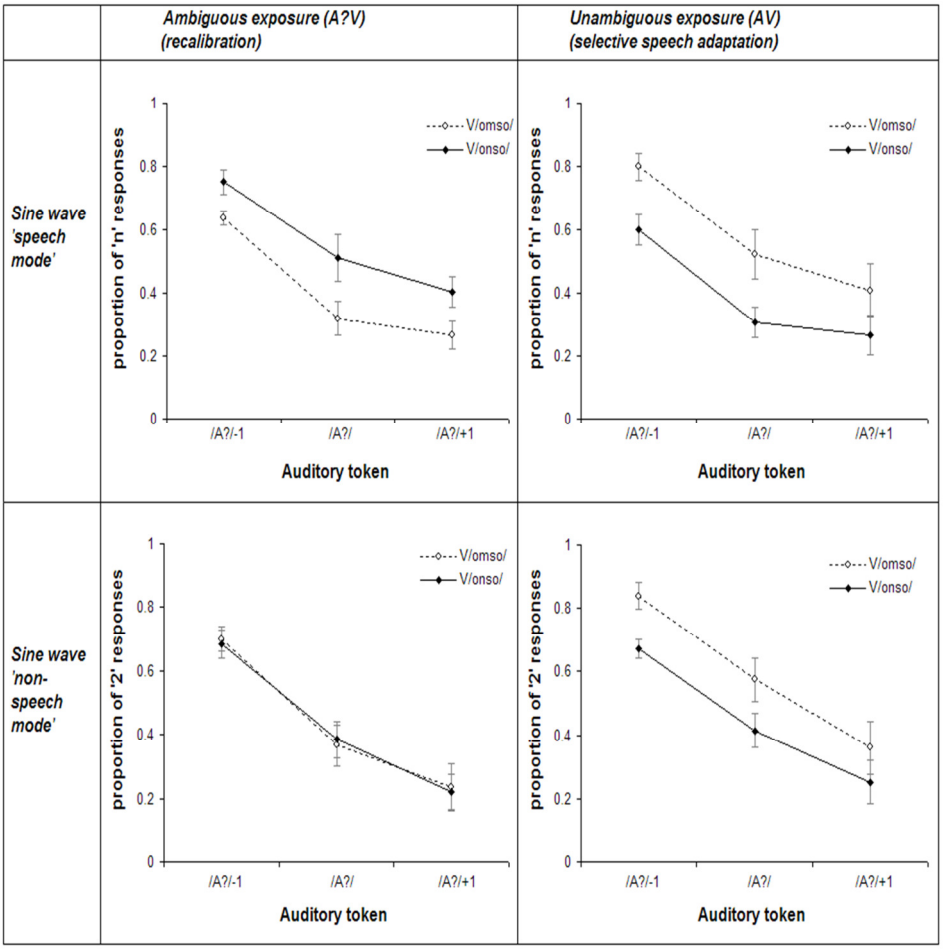

Figure 3c. *Mean proportion of /onso/ (or '2') responses as a function of the auditory test tokens of the continuum after exposure to auditory ambiguous adapters A?Vonso and A?Vomso (left panels), and auditory non-ambiguous adapters AonsoVonso and AomsoVomso (right panels). The upper panels show performance of the speech group, the lower panels of the non-speech group. Error bars represent one standard error of the mean.*

Mode	Exposure sound	Lipread adapter		Aftereffect
		/onso/	/omso/	
Speech Mode	Ambiguous	.55	.41	.14*
	Non-ambiguous	.39	.58	-.19*
Non-speech Mode	Ambiguous	.43	.43	.00
	Non-ambiguous	.46	.59	-.13*

* Significance at $p < .05$

Table 3.2. *Mean proportion of 'onso'- or '2'-responses and the corresponding aftereffect after exposure to audiovisual adapters with ambiguous and nonambiguous sounds.*

3.5 - Discussion

The present results clearly demonstrate that recalibration of phonetic categories and selective speech adaptation can be obtained with sine wave replicas. Moreover, the use of SWS allowed us to observe a remarkable dissociation between these two phenomena: recalibration was observed only when listeners were in speech mode, whereas selective adaptation occurred for listeners in speech and non-speech mode. Previous studies already demonstrated that these two phenomena not only differ in the direction of their aftereffect, but also in the speed with which they build-up and dissipate (recalibration builds up fast and peaks early, selective adaptation builds up slowly and increases with prolonged exposure (Vroomen et al., 2004, 2007)). Together, these dissociations therefore provide strong evidence that there are distinct mechanisms underlying recalibration and selective adaptation.

Our findings on selective speech adaptation fit well with previous reports showing that low-level mechanism are mainly responsible for the effect to occur. For example, Roberts and Summerfield (1981) demonstrated that adaptation was induced by the auditory component, whereas the phonetic label attached to the adapting stimulus had no effect. Here we also observed that equal amounts of selective adaptation were obtained for listeners in speech or non-speech mode. This again suggests that it is the acoustic and non-ambiguous nature of the adapter that causes selective adaptation, while the more high-level interpretation of the stimulus has little or no effect. In that sense, adaptation is also similar to other forms of perception like color, curvature (Gibson, 1933) or motion (Anstis, 1986, chapter 16) where aftereffects mainly depend on the non-ambiguous visual nature of the adapting stimulus.

In stark contrast with selective adaptation, recalibration appeared to be speech-specific. The notion underlying recalibration is that reliable information from one source disambiguates unreliable information from another source. Here, it was lipread speech that provided reliable information about how to interpret an ambiguous 'm/n' sound. Presumably, during exposure there is a conflict between the heard and lipread information that is resolved by shifting the phoneme boundary so that the ambiguous sound matches the lipread information. This shift occurs quite fast (Vroomen et al., 2007) and it lasts for some time so that it is observable as an aftereffect. It seems only logical that recalibration occurs to the extent that the conflicting information sources are referring to the same distal event, here whether the speaker said /m/ or /n/. For listeners in speech mode, both inputs were indeed combined into a single phonetic presentation as observable in a direct bias effect on the goodness rating of the sound. Listeners in non-speech mode, though, were not under the impression that the auditory and visual signal referred to the same event, and the two information streams were therefore treated as separate. Listeners labeling the SWS sounds as '1' or '2' thus made no connection with the segmental content of the simultaneously presented lipread information, and there was therefore also no effect of lipreading on the goodness ratings of the SWS sound if perceived as non-speech.

The use of the SWS stimuli to induce recalibration and selective speech adaptation may also provide new opportunities to explore the nature of these phenomena. Eisner and McQueen (2005) reported that recalibration for the fricatives (/s/–/f/) did not generalize to a novel speaker. Similarly, Kraljic and Samuel (2005) tested the fricatives (/s/–/S/) and found that tuning did not generalize across speakers. When a male voice was heard during the exposure phase, at test recalibration was reliable for male-produced tokens but not for female-produced tokens. Kraljic and Samuel (2006) also tested stop consonants (/d/–/t/) and here they did observe that recalibration generalized to a novel speaker. They argued that the patterns of generalization may be due to the acoustic similarity among the different exposure and test tokens. On this view, recalibration generalizes to acoustically similar sounds, but not to acoustically dissimilar sounds (see also Mirman et al., 2006). It remains for future studies to explore whether there is generalization from SWS sounds to natural speech and vice versa, and whether the same holds for selective speech adaptation. If it is indeed the acoustic similarity across tokens that determines whether recalibration will generalize, one may find that there is no generalization from SWS tokens to natural speech, while there is generalization for selective speech adaptation.

Chapter 4

Lipreading recalibrated by speech sounds [4]

[4]Adapted from:

Baart, M. & Vroomen, J. (2010a). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters, 471*, 100-103.

4.1 - Abstract

It is well known that visual information derived from mouth movements (i.e., lipreading) can have profound effects on auditory speech identification (e.g., the McGurk-effect). Here we examined the reverse phenomenon, namely whether auditory speech affects lipreading. We report that speech sounds dubbed onto lipread speech affect immediate identification of lipread tokens. This effect likely reflects genuine cross-modal integration of sensory signals and not just a simple response bias because we also observed adaptive shifts in visual identification of the ambiguous lipread tokens after exposure to incongruent audiovisual adapter stimuli. Presumably, listeners had learned to label the lipread stimulus in accordance with the sound, thus demonstrating that the interaction between hearing and lipreading is genuinely bi-directional.

4.2 - Experiment 1, Introduction

The question of how sensory modalities cooperate in forming a coherent representation of the environment is the focus of much current work. A particularly elucidating example is the interaction

between hearing and seeing speech (here referred to as lipreading). In one of the more spectacular cases, listeners report to 'hear' /da/ when in fact, auditory /ba/ is dubbed onto lipread /ga/, the McGurk-effect (McGurk & MacDonald, 1976). Numerous studies have explored the brain mechanisms underlying this phenomenon. Some have reported that visual speech may affect auditory processing as early as the auditory cortex (Calvert et al., 1997; Colin et al., 2002; Möttönen, Krause, Tiippana, & Sams, 2002; Pekkola et al., 2005; Sams et al., 1991). The interaction has been found to occur between 150 and 250 ms using the mismatch negativity paradigm (Colin et al., 2002; Möttönen et al., 2002; Sams et al., 1991), while others have reported that as early as 100 ms, the auditory N1 component is attenuated and speeded up when auditory speech is accompanied by lipread information (Besle et al., 2004; van Wassenhove et al., 2005), possibly because visual speech predicts when a sound is going to occur (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010).

Notably though, to date it is not known whether auditory speech also affects visual processing of lipread speech. This is surprising because bi-directional effects have been reported in other crossmodal illusions. For example, in the 'ventriloquist illusion', the apparent location of a sound is displaced towards a simultaneously presented and spatially misaligned light (Bertelson, 1999). The reverse phenomenon, namely that the apparent location of a visual target is shifted towards an auditory displaced distracter, has also been reported (Radeau & Bertelson, 1987), although the effect is admittedly small because the more reliable information source, - for space vision - is dominant and thus less susceptible to cross-modal biases (Ernst & Bülthoff, 2004; Hidaka et al., 2009).

In a first experiment, we sought to show that identification of lipread stimuli is affected by speech sounds. For that purpose, we created a 7-point continuum of visual stimuli in between /omso/ and /onso/. Participants were instructed to lipread these stimuli and press an 'm'- or 'n'-key upon lipreading /omso/ or /onso/, respectively (a visual 2AFC-task), while trying to ignore /omso/ or /onso/ sounds that were dubbed onto the videos. Despite instructions to ignore the sound, we expected the sound to shift the visual identification function of the lipread stimuli, so more 'n'-responses upon hearing /onso/ rather than /omso/.

4.3 - Experiment 1, Method

4.3.1 - Participants

Twelve native speakers of Dutch (mean age = 23) with normal hearing and normal or corrected-to-normal vision participated after giving written informed consent. The experiment was conducted in accordance with the Declaration of Helsinki.

4.3.2 - Stimuli

Stimulus creation started with two videos of the full face of a male speaker pronouncing the pseudo-words /omso/ and /onso/ as previously used by Tuomainen et al. (2005). The head, nose, and eye position of the speaker were well aligned, so fusion of the stimuli could be accomplished by adjusting the overall opacity rather than applying a morphing technique with landmarks on the face. The lipread /m/ and /n/ belong to different 'viseme' classes (Walden, Prosek, Montgomery, Scherr, & Jones, 1977), and are thus relatively easy to visually discriminate. To create a continuum in between the two recordings, videos were first converted into bitmap sequences (29.97 f/s) matched for total duration (45 bitmaps; 1500 ms) and for onset and offset of the articulatory gesture (at 567 and 1367 ms, respectively). Each individual bitmap of the /omso/ sequence was fused with the corresponding /onso/ bitmap by adding the two bitmaps in different relative proportions to each other. Seven bitmap sequences were created by varying the relative proportion from 0 to 100% for the most /omso/-like stimulus, through 15-85%, 29-81%, 43-57%, 58-42%, 72-28%, to 90-10% for the most /onso/-like stimulus. Each of the seven thus created videos (14.9 (H) by 18.8 (W) cm in size) looked natural without any noticeable jitter or fading. The natural timing between the audio and video was preserved by relying on a custom-made program that displayed the bitmap sequence and played the sound by trigger, rather than a standard PC-based video-player whose timing was considered to be too unreliable.

4.3.3 - Procedure and design

Participants were tested individually in a sound attenuated and dimly lit booth and were seated at approximately 70 cm from a 17-inch CRT screen. The audio was presented at 63 dBa at ear-level via two regular loudspeakers placed left and right of the monitor. The seven lipread tokens of the continuum were delivered in combination with auditory /omso/ and /onso/, and in a silent condition. Each of the 21 stimuli was delivered 20 times (420 trials in total) in four blocks of 105 randomly presented trials. Participants judged whether the visual stimulus was /omso/ or /onso/ by pressing the corresponding 'm'- or 'n'-key. The next trial started 750 ms after a response was detected. Prior to testing, participants received a short practice session (12 trials) in

which they were shown the two extreme tokens of the lipread continuum combined with auditory /omso/, /onso/, or silence. It was stressed that participants had to rely on lipreading rather than sound, as the sound did not predict in any sense what the visual stimulus would look like.

4.4 - Experiment 1, Results and discussion

For each participant, the proportion of 'n'-responses was determined as a function of the lipread token. The group-averaged data are presented in Figure 4a. As is clearly visible, three rather sharp S-shaped visual identification functions were obtained. The 50% cross-over point of the curves was near the middle of the continuum and the

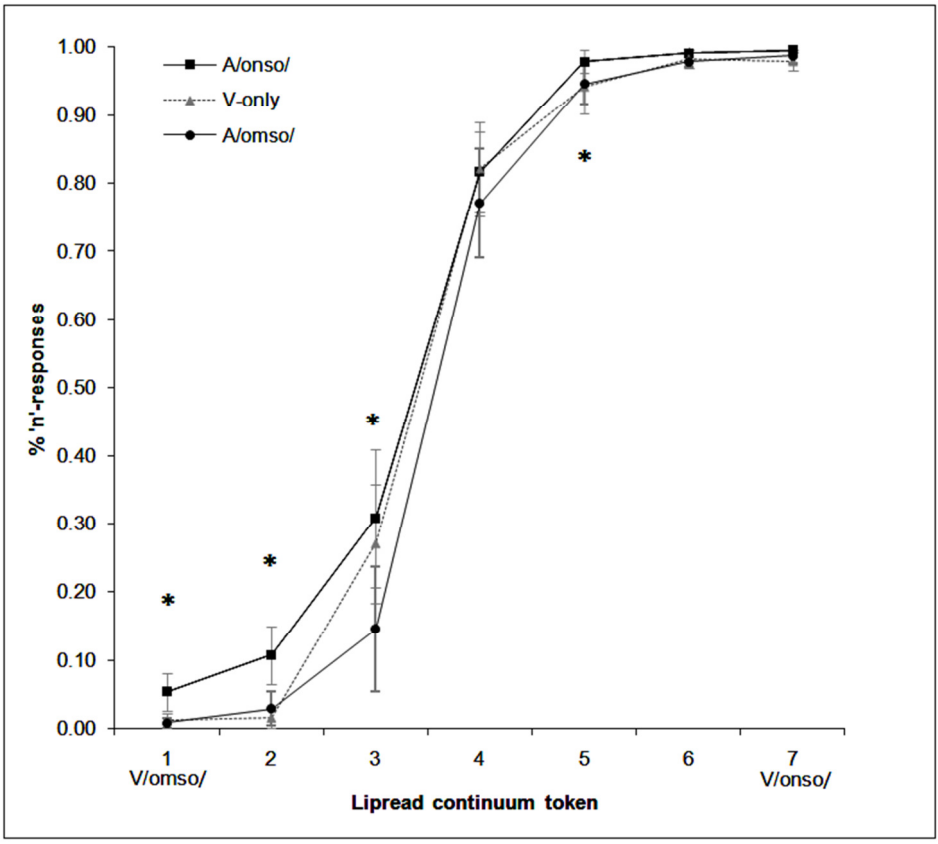

Figure 4a. *The proportion of 'n'-responses as a function of the lipread token of the continuum, separately for the visual-only condition, and when combined with auditory /onso/ or /omso/. Significant differences between individual lipread tokens combined with auditory /onso/ versus /omso/ are denoted by an asterisk. Error bars represent one standard error of the mean.*

extremes were almost entirely judged as /omso/ or /onso/. This demonstrates that our continuum was adequately created. Most importantly, the dubbing of a sound onto the videos shifted the visual identification functions in the predicted direction, so more 'n'-responses if /onso/ was dubbed onto the video rather than /omso/. A 3 (Aonso, Aomso, or V-only) x 7 (lipread token) ANOVA on the proportion of 'n'-responses showed a main effect of lipread token ($F(6,66) = 180.73$, $p < .001$) because - unsurprisingly - there were more 'n'-responses if the lipread videos contained a larger portion of the original /onso/-video. Most importantly, there was a main effect of sound ($F(2,22) = 4.61$, $p < .022$) because there were more 'n'-responses if auditory /onso/ was dubbed onto the video rather than /omso/. The interaction was also significant ($F(12,132) = 2.08$, $p < .023$). Separate paired t-tests confirmed that there were more 'n'-responses on the first, second, third, and fifth lipread token if that token was combined with auditory /onso/ rather than /omso/ (all p's$_{\text{one-tailed}} < .05$).

These data thus clearly demonstrate that a speech sound does indeed affect lipreading. The question posed in the introduction, namely whether the cross-modal interaction between speech and lipreading is bi-directional can thus, as a first approximation, be answered affirmatively. However, a critical issue is to determine the processing stage at which this effect occurs. At least two possibilities are available. On the one hand, it may be that the auditory-induced shift is reflecting a truly perceptual effect of sound on vision. Alternatively, though, it might also reflect a response strategy of the participant who, whenever unsure about the visual target, relied on the sound that was heard, despite instructions to ignore that sound.

4.5 - Experiment 2, Introduction

To further examine this, we conducted another experiment in which we measured aftereffects using an exposure-test paradigm as introduced by Bertelson et al. (2003). In that study, it was reported that if an ambiguous sound halfway between /b/ and /d/ was dubbed onto lipread /b/ (rather than /d/), participants were more likely to categorize the initially ambiguous sound as /b/ when tested later in an auditory-only speech identification test. Presumably, listeners had learned to label the ambiguous sound in accord with the lipread information (i.e., phonetic recalibration). This finding was taken as a particularly clear example that lipreading affects speech identification beyond the level of simple response biases. This finding has been replicated in many other studies (e.g., van Linden & Vroomen, 2007; Vroomen & Baart, 2009a, 2009b; Vroomen et al., 2007; Vroomen et al., 2004). Here, we tested the reverse situation, namely whether a sound would induce a longer-lasting change about the interpretation of an initially ambiguous lipread stimulus. To ensure that this was not due to response

priming (i.e., respond /onso/ during test if /onso/ was presented in foregoing exposure phase) we included, as in Bertelson et al. (2003), a control condition with stimuli that were not expected to induce recalibration. For that purpose, we used the extreme video tokens /omso/ or /onso/ with the congruent sounds (VmAm and VnAn) dubbed onto them. These stimuli were not expected to induce recalibration because there is no deviance between sight and sound that supposedly drives recalibration. However, the unambiguous nature of the lipread stimuli might possibly cause a contrastive aftereffect (in the auditory domain known as 'selective speech adaptation') as has been demonstrated before for auditory speech (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994), color, curvature (Gibson, 1933), or motion (Anstis, 1986), possibly reflecting a 'fatigue' of some hypothetical feature detectors. With unambiguous audiovisual exposure stimuli, one might thus expect fewer 'n'-responses after exposure to VnAn than VmAm, an effect in the opposite direction of recalibration.

4.6 - Experiment 2, Method

4.6.1 - Participants

Twenty-two new native speakers of Dutch with normal hearing and normal or corrected-to-normal vision participated (mean age = 21).

4.6.2 - Procedure and design

A pre-test was used to determine the most ambiguous lipread token for each participant. Each video of the continuum was delivered 16 times in random order and participants indicated whether they saw /omso/ or /onso/. The video closest to the individually determined 50% cross-over point was taken as the perceptually most ambiguous video (henceforth V?).

During adaptation, participants were repeatedly exposed to a short block of audiovisual adapter stimuli and then tested on lipreading. Each exposure block contained eight consecutive presentations (ISI = 500 ms) of one of the four audiovisual adapter stimuli V?An or V?Am (to induce recalibration), or VnAn or VmAm (to induce selective adaptation). Exposure was immediately followed by a lipreading test consisting of three different videos presented twice in random order (six test trials in total). The three videos were V?, its immediate 'omso-like' neighbour on the continuum V? − 1 and its immediate 'onso-like' neighbour V? + 1. During the test, participants indicated whether they saw the speaker pronounce /onso/ or /omso/ by pressing a corresponding key.

There were 32 exposure-test blocks in total (8 for each of the 4 adapters), delivered in pseudo-random order. At the end of the experiment, participants were also

asked to rate the /omso/–/onso/ quality of the lipread part of the audiovisual adapter stimuli on a seven point Likert-scale with '1' representing a clear visual /omso/ and '7' a clear visual /onso/. Each of the four adapters was presented six times (ISI = 900 ms) in pseudo-random order.

4.7 - Experiment 2, Results

The number of 'n'-responses was determined for each participant (see Figure 4b for the group averages) and these data were submitted to a 2 (ambiguous or unambiguous lipread exposure) x 2 (adapter sound /omso/ or /onso/) x 3 (lipread test-token) overall ANOVA. There was a main effect of ambiguity of the lipread adapter ($F(1,21) = 5.86$, $p < .025$) because there were somewhat more 'n'-responses after exposure to the ambiguous adapters than the unambiguous ones. The main effect of the lipread testtoken ($F(2,42) = 116.42$, $p < .001$) indicated that there were more 'n'-responses for the more 'onso-like' token (V? + 1) than for V? and V? − 1. Most importantly, there was an interaction between ambiguity of the lipread adapter and identity of the adapter sound ($F(1,21) = 12.55$, $p < .002$) indicating that there were *more* 'n'-responses after exposure to V?An than V?Am (recalibration), but *fewer* 'n'-responses after exposure to VnAn than VmAm (selective adaptation).

To isolate these effects, aftereffects were calculated analogous to previous studies by subtracting the proportion of 'n'-responses after exposure to auditory /omso/ from /onso/, thereby pooling over the three test-tokens (see Table 4.1).

Separate t-tests showed that, in total, there were 3% more 'n'-responses after exposure to V?An than V?Am, $t(21) = 2.10$, $p_{\text{one-tailed}} < .024$, while there were 5% less 'n'-responses after exposure to VnAn than VmAm, $t(21) = 2.66$, $p_{\text{one-tailed}} < .008$. Figure 4b shows that the aftereffect was mainly restricted to the most ambiguous token V?. A separate t-test isolating performance on the ambiguous V? token showed that there were 6% more 'n'-responses after exposure to V?An than V?Am, $t(21) = 1.91$, $p_{\text{one-tailed}} < .035$, while there were 12% fewer 'n'- responses after exposure to VnAn than VmAm, $t(21) = 3.39$, $p_{\text{one-tailed}} < .002$. Thus, as predicted, a learning effect was observed if the audiovisual adapter contained an ambiguous lipread token, and a contrast effect if it contained an unambiguous lipread token.

The goodness ratings about the visual part of the audiovisual adapters further confirmed that the ambiguous token V? was rated more 'onso'-like if combined with auditory /onso/ rather than /omso/ (4.51 versus 3.00 on a 7-point scale, respectively, $t(21) = 3.62$, $p_{\text{one-tailed}} < .001$). We also tested the possibility that participants who showed bigger learning effects were also more influenced by the sound of the adapter. The effect of the sound was calculated by taking the difference between the goodness

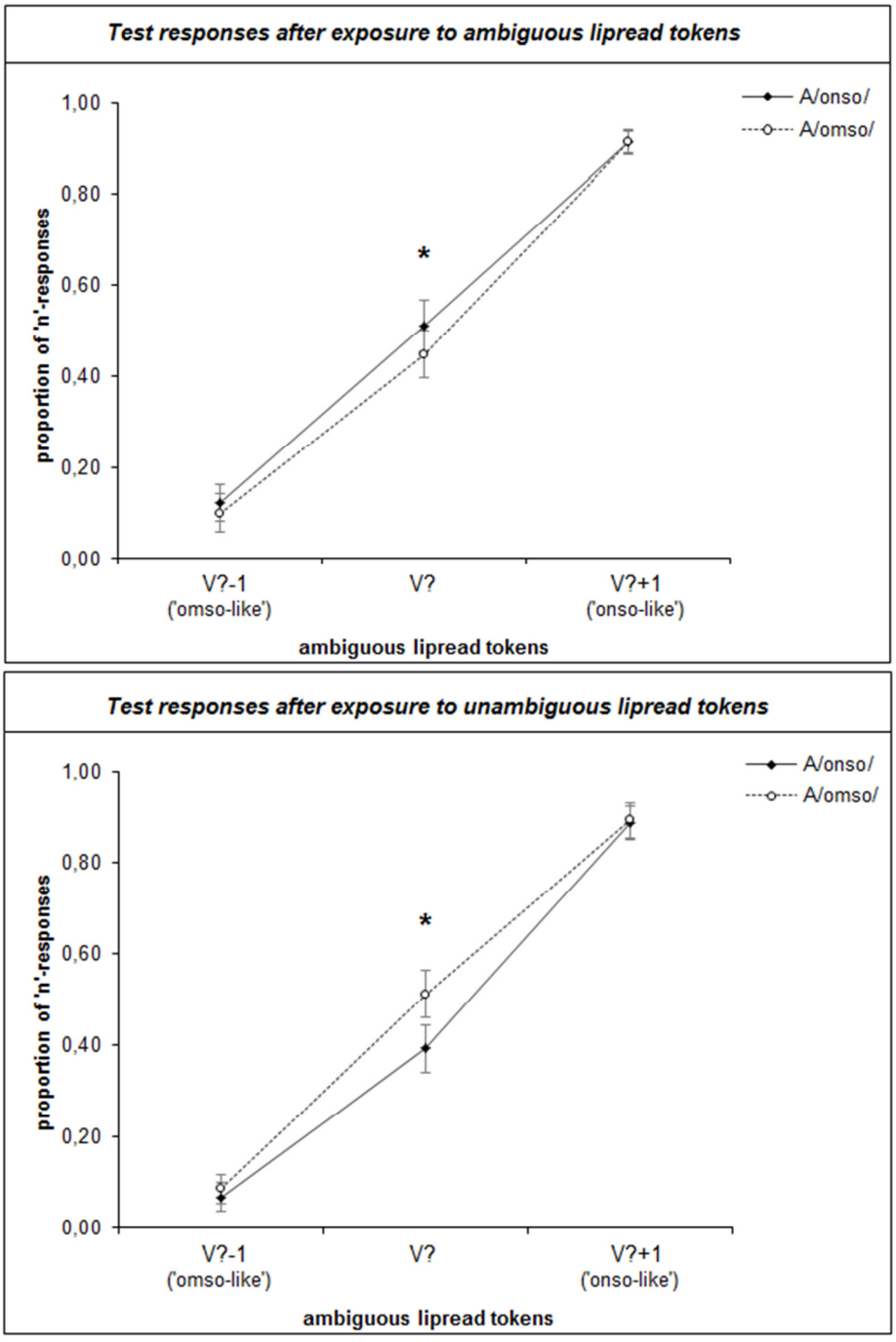

Figure 4b. *The mean proportion of 'n'-responses on lipread test-tokens after exposure to ambiguous lipread adapters (AnV? and AmV?; upper panel), and unambiguous lipread adapters (AnVn and AmVm; lower panel). Significant differences between test-tokens preceded by exposure to auditory /onso/ versus /omso/ are denoted by an asterisk. Error bars represent one standard error of the mean.*

Lipread information	Auditory information		Aftereffect
	/onso/	/omso/	
Ambiguous (V?)	.52	.49	.03
Non-ambiguous (Vn or Vm)	.45	.50	-.05

Table 4.1. *Mean proportion of 'n'-responses and the corresponding aftereffect after exposure to audiovisual adapters with ambiguous and unambiguous lipread videos.*

rating of V?An and V?Am, and this difference indeed correlated with the size of the recalibration effect (r = .39, $p_{\text{one-tailed}} < .038$). Participants who were strongly affected by the sound thus displayed larger recalibration effects at test.

4.8 - General discussion

The present study thus clearly demonstrated that an ambiguous lipread stimulus between /m/ and /n/ is more likely labelled as 'n' if a /n/-sound is dubbed onto it rather than /m/. This immediate effect is not due to a response bias only because we also observed a longer-lasting learning effect: that is, the ambiguous video was labelled more likely as 'n' if in a preceding adapter phase an /n/-sound was dubbed onto it rather than /m/. Presumably, exposure to the audiovisual adapter resulted in an endurable adjustment of the boundary of the ambiguous lipread token that - in later testing - was still observable as an aftereffect. For ambiguous lipread tokens, participants thus adjust the phoneme boundary such that the conflict between heard and lipread information is reduced. These findings are in close correspondence with previous reports on phonetic adjustments in auditory speech (e.g., Bertelson et al., 2003; van Linden & Vroomen, 2007; Vroomen & Baart, 2009a; Vroomen et al., 2007; Vroomen et al., 2004), thus indicating that similar mechanisms underlie auditory and lipread recalibration. Moreover, simple response priming (e.g., respond 'n' at test if previously exposed to /onso/) can also be excluded as a mechanism that accounts for these effects because unambiguous and audiovisual congruent adapters produced *contrastive* aftereffects. These visual contrast-effects have been demonstrated before for auditory speech, color, curvature, and so forth, but here we provide the first demonstration of their occurrence for lipread speech.

What might be the functional reason that there is an interaction between seeing and hearing speech? At least two relevant notions have appeared in the literature. The first is that it is 'ecologically' useful to consult more than one source, primarily because

different sense organs provide complementary information about the same external event. For this reason, lipreading is used in understanding speech as it can compensate for interference from external noise and may resolve internal ambiguities of the auditory speech signal. A second reason is that there is internal 'drift' or 'error' within the individual senses that can be adjusted by cross-reference to other modalities. In the spatial domain of sensor-motor adaptation to optical-wedge prisms, this is already known for more than 100 years (von Helmholz, 1866), but for speech, this kind of cross-reference to other modalities has been reported only very recently (Bertelson et al., 2003). In both cases, though, there is a perceptual adjustment induced by a deviance between two information sources that the brain tries to reduce. The present study extends these findings by showing that this kind of adjustment not only occurs for auditory, but also for visual speech.

Our findings are also of relevance for the neural mechanisms involved in multisensory processing of audiovisual speech. Neuroimaging and electrophysiological studies have found audiovisual interactions in multimodal areas such as the superior temporal sulcus (STS) and sensory-specific areas including the auditory and visual cortices (Besle et al., 2004; Callan et al., 2004). It has been proposed that the unimodal inputs are initially integrated in STS and that interactions in the primary auditory and visual cortices reflect feedback from STS (Calvert et al., 1999). On this account, interactions in the primary cortex are presumably mediated by the STS via backward projections (Besle et al., 2004). Besides STS, motor regions of planning and execution (Broca's area, premotor cortex, and anterior insula) could be involved via the so-called mirror neurons (e.g., Giard & Peronnet, 1999; Klucharev et al., 2003; Ojanen et al., 2005; Skipper, Nusbaum, & Small, 2005). Broca's area is proposed to be a homologue of the macaque inferior premotor cortex (area F5) where mirror neurons are situated that discharge upon action and perception of goal-directed hand or mouth movements. The presumed function of these mirror neurons is to mediate imitation and aid action and understanding (Rizzolatti & Craighero, 2004). Broca's area is not only involved in the production of speech, but is also activated during silent lipreading (Campbell et al., 2001) and passive listening to speech (Wilson, Saygin, Sereno, & Iacoboni, 2004). On this view, activation of mirror neurons in Broca's area may facilitate a link between auditory and visual speech inputs and the corresponding motor representations. In line with this notion, it has been reported that recalibration of auditory 'sine-wave speech' by lipread information occurs only if the sine-wave tokens were perceived as speech, but not if they were perceived as non-speech sounds (Vroomen & Baart, 2009a), most likely because in the latter case, there was no link to articulatory motor programs. Vision may thus affect auditory processing via articulatory motor programs of the

observed speech acts (Callan et al., 2003), and as demonstrated here, it is conceivable that this effect is bi-directional in nature.

Chapter 5

Phonetic recalibration and working memory [5]

[5]Adapted from:

Baart, M. & Vroomen, J. (2010b). Phonetic recalibration does not depend on working memory. *Experimental Brain Research, 203,* 575 - 582.

5.1 - Abstract

Listeners use lipread information to adjust the phonetic boundary between two speech categories (phonetic recalibration, Bertelson et al. 2003). Here, we examined phonetic recalibration while listeners were engaged in a visuospatial or verbal memory working memory task under different memory load conditions. Phonetic recalibration was - like selective speech adaptation - not affected by a concurrent verbal or visuospatial memory task. This result indicates that phonetic recalibration is a low-level process not critically depending on processes used in verbal- or visuospatial working memory.

5.2 - Introduction

In natural speech, there are other information sources besides the auditory signal that facilitate perception of the spoken message. For example, viewing a speaker's articulatory movements (i.e., lipreading) is known to improve auditory speech intelligibility (e.g., Erber, 1974), especially when the auditory input is ambiguous (Sumby & Pollack, 1954). More recent work has demonstrated that listeners also use lipread information to adjust the phonetic boundary between two speech categories (Bertelson et al. 2003; Vroomen et al. 2004, 2007; van Linden and Vroomen 2007, 2008; Vroomen and Baart 2009b). For example, listeners exposed to an auditory ambiguous speech sound halfway between /b/ and /d/ (i.e., A? for auditory ambiguous) that is combined with the video of a speaker articulating either /b/ or /d/ (Vb and Vd for visual /b/ or /d/, respectively) report, in a subsequently delivered auditory-only test, more 'b'-responses after exposure to A?Vb than after A?Vd, as if they had learned to label the ambiguous sound in accordance with the lipread information (i.e., phonetic recalibration). Lipread-induced recalibration of phonetic categories has now been demonstrated many times (Vroomen et al., 2004, 2007; van Linden & Vroomen, 2007, 2008; Vroomen & Baart, 2009a, b) and has also been demonstrated to occur if the disambiguating information stems from lexical knowledge about the possible words in the language rather than from lipread information (e.g., Norris et al., 2003; Kraljic & Samuel, 2005, 2006, 2007; van Linden & Vroomen, 2007).

The mechanism underlying phonetic recalibration though, is at present largely unknown. A recent functional magnetic resonance imaging (fMRI) study (Kilian-Hütten, Vroomen, & Formisano, 2008) using the same stimuli and design as in Bertelson et al. (2003) showed that the trial-by-trial variation in the amount of recalibration could be predicted from activation in the middle/inferior frontal gyrus (MFG/IFG) and the inferior parietal cortex. These brain areas are also known to be involved in verbal working memory (Jonides et al., 1998), and it might thus be conceivable that phonetic recalibration shares neural underpinnings with verbal working memory. Alternatively, though, there is behavioral and neurophysiological evidence which shows that lipreading has profound effects on speech perception at very early processing levels and that the effect is quite automatic (Colin et al., 2002; Massaro, 1987, 1998; McGurk & MacDonald, 1976; Möttönen et al., 2002; Soto-Faraco, Navarra, & Alsius, 2004). On this view, it may seem more likely that lipread-induced recalibration would not rely on high-level neural resources used for working memory, because it is basically a low-level process operating in an automatic fashion.

To examine whether phonetic recalibration and working memory indeed share common resources, we measured phonetic recalibration while participants were engaged

in a working memory task. In the literature on working memory, a distinction is usually made between a verbal and a visuospatial component (e.g., Baddeley & Hitch, 1974; Baddeley & Logie, 1999), which rely on distinct neural structures. For example, Smith, Jonides and Koeppe (1996) showed primarily left-hemisphere activation during a verbal memory task, whereas the visuospatial task mainly activated right-hemisphere regions.

As a control for general disturbances caused by the dual task, we also examined whether the verbal and spatial memory task would interfere with selective speech adaptation. Selective speech adaptation, first demonstrated by Eimas and Corbit (1973), depends on the repeated presentation of a particular speech sound that causes a reduction in the frequency with which that token is reported in subsequent identification trials. Since its introduction, many questions have been raised about the nature underlying this effect. Originally, it was thought to reflect a fatigue of some hypothetical 'linguistic feature detectors', but others argued that it reflects a shift in criterion (e.g., Diehl et al., 1978), or a combination of both (Samuel, 1986). Others (e.g., Ganong, 1978) however, have argued that the size of selective speech adaptation depends upon the amount of spectral overlap between adapter and test sound. As such, most of the effect would then be auditory rather than phonetic in nature. Moreover, selective speech adaptation is automatic as it is unaffected by a secondary on-line arithmetic or rhyming task (Samuel & Kat, 1998). Following this line of reasoning, we did not expect our working memory task to interfere with selective speech adaptation.

To induce phonetic recalibration and selective speech adaptation, we used the same stimuli and procedures as in Bertelson et al. (2003). Participants were presented with multiple short blocks of eight audiovisual exposure trials immediately followed by six auditory-only test trials. During each exposure-test block, participants tried to memorize a set of previously presented letters for the verbal memory task or a motion path of a moving dot for the spatial task. The difficulty of the secondary memory task was increased across three groups of participants up until the point that performance on both memory tasks was about equal, sufficiently above chance level but below ceiling.

To the extent that phonetic recalibration shares mechanisms with working memory, one might expect more interference from the verbal rather than spatial memory task because lipreading also relies primarily on activation in the left hemisphere (Calvert & Campbell, 2003). Moreover, interference should increase if the memory task becomes more demanding. Alternatively, though, if recalibration is, like selective speech adaptation, a low-level process running in an automatic fashion, then neither the verbal nor the spatial memory task should interfere with recalibration.

5.3 - Method

5.3.1 - Participants

Sixty-six native speakers of Dutch (mean age = 21 years) with normal hearing and normal/corrected to normal vision participated, twenty-two in each of three memory load conditions. All participants gave their written informed consent prior to testing, and the experiment was conducted according to the Declaration of Helsinki.

5.3.2 - Stimuli

The audiovisual adapter stimuli are described in detail in Bertelson et al. (2003). In short, the audio tracks of audiovisual recordings of a male speaker of Dutch pronouncing /aba/ and /ada/ were synthesized into a nine-step /aba/-/ada/ continuum in equal Mel-steps. To induce recalibration, the token from the middle of the continuum (A?) was dubbed onto both videos so as to create A?Vb and A?Vd. To induce selective speech adaptation, two audiovisual congruent adapters were created by dubbing the continuum endpoints onto the corresponding videos for AbVb and AdVd. As test stimuli served the most ambiguous sound on the continuum /A?/ and its immediate continuum neighbors /A?-1/ (more '/aba/-like') and /A?+1/ (more '/ada/-like').

5.3.3 - Design and procedure

Participants were tested individually in a sound-attenuated and dimly lit booth. They sat at approximately 70 cm from a 17-inch CRT screen. The audio was delivered via two regular loudspeakers placed left and right of the monitor at 63 dBa (measured at ear level). The videos showed the speaker's entire face from the throat up to the forehead and were presented against a black background in the center of the screen (W: 10.4 cm, H: 8.3 cm). Testing was spread out over two subsequent days. Half of the participants were tested for recalibration on the first day, and selective speech adaptation on the second day, for the other half of the participants the order was reversed. On both days, participants were tested in three separate blocks. One was a single-task adaptation procedure that served as baseline, the others were dual-task procedures using a visuospatial or a verbal memory task. Block order was counterbalanced across participants in a Latin square.

5.3.3.1 - Recalibration/selective adaptation procedure

To induce recalibration, participants were exposed to eight repetitions (ISI = 425 ms) of either A?Vb or A?Vd. The exposure phase was immediately followed by an auditory-only test containing the ambiguous test stimulus /A?/, and its immediate neighbors on the continuum /A?-1/ and /A?+1/. These three test stimuli were presented

twice in random order. After each test trial, participants had to indicate whether they heard /aba/ or /ada/ by pressing the corresponding 'b'- or 'd'-key on a response box. The next test trial was delivered 1,000 ms after a key press. There were sixteen exposure-test blocks (eight for A?Vb, and eight for A?Vd), delivered in pseudo-random order.

The procedure to induce selective speech adaptation was exactly the same as for recalibration, except that participants were exposed to AbVb and AdVd. To ensure that participants attended the lipread videos during exposure, they were instructed - as in previous studies - to indicate whether they noticed an occasional small white dot on the upper lip of the speaker (12 px in size, 120 ms in duration).

5.3.3.2 - Working memory tasks

In an attempt to equate task difficulty of the verbal and visuospatial memory tasks, we had to manipulate the set size of the memory items in a non-symmetrical way. Verbal items were easier to remember than the visuospatial ones and for this reason, the number of memory items in both tasks differed as specified below.

5.3.3.3 - The visuospatial task

For the visuospatial task, each exposure-test block was preceded by a newly generated random path of a white dot (Ø = .4 cm) that moved across a dark screen in three (for the low-memory load group) or four (for the intermediate and high-memory load groups) steps. Each dot was presented for 500 ms. Participants were instructed to carefully attend to the target path and to remember it by covert repetition throughout the entire exposure-test block that would follow the target path. The exposure-test block was delivered to induce and measure recalibration or selective speech adaptation 1,300 ms after the last dot had disappeared. Immediately after this exposure-test block, participants were then presented a spatial probe for which they indicated whether its motion path was the same or different as the target by pressing a 'yes'- or 'no'-key (see Figure 5a(A)). In half of the trials, the target and the probe were the same, in the other half of the trials, the probe differed by one dot.

5.3.3.4 - The verbal memory task

For the verbal memory task, participants had to remember a string of three (the low-memory load group), five (the intermediate-memory load group) or seven (the high-memory load group) letters that appeared simultaneously in the center of the screen for 2,000 ms. Participants were instructed to covertly repeat the string of letters throughout the exposure-test block that would follow. After the exposure - test block,

a one-letter test probe was presented for which participants indicated whether it was one of the targets by pressing the 'yes'- or 'no'-key (Fig. 5a(B)). Half of the trials required a 'yes'-response. The target letters were chosen from 16 consonants of the Latin alphabet, excluding 'B' and 'D', because they made up the crucial phonetic contrast. All letters were displayed in capitals (font type: Arial; size: 1.3(W) by 1.6(H) cm; spacing: 2.0 cm).

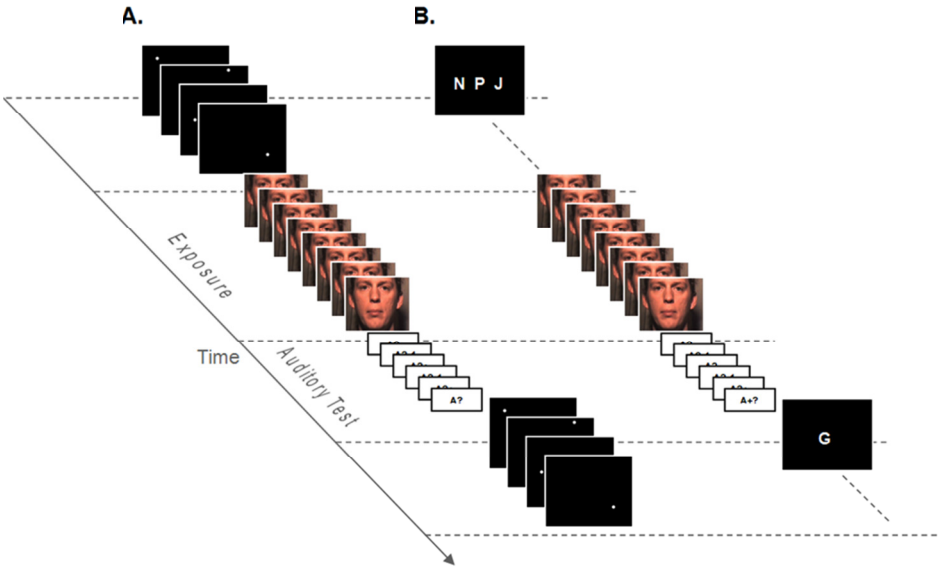

Figure 5a. *Schematic overview of an exposure-test block in the low-load memory condition. In the visuospatial memory task (A), the motion path of a dot had to be remembered during the audiovisual exposure - auditory-only test phase. The memory probe immediately followed the final test token. In the verbal task (B), three letters had to be remembered.*

5.4 - Results

5.4.1 - Performance on the memory tasks

The average number of correct responses in the verbal and spatial memory task under the three load conditions is presented in Table 5.1. In the ANOVA on the percentage of correct responses, the main effect of task, $F(1,64) = 40.40$, $p < .001$, showed that verbal probes were recognized somewhat better than the spatial probes, (91 vs. 82%, respectively, with chance level at 50%). There was also a main effect of load, $F(1,64) = 23.30$, $p < .001$, because recognition became worse when load increased. There was an interaction between memory load and task; $F(1,64) = 15.24$, $p < .001$, as increasing the memory load had a bigger impact on the verbal task (where set size was increased from 3 to 7 items) than the spatial task (where the target path was increased from 3 to 4 steps from low to medium, and remained at 4 during high load). As

intended, in the high-load condition, overall performance for the verbal and spatial task were not different from each other (p = .88), so task difficulty was equated here. The results for the memory task confirm that participants were indeed paying attention to the task as performance was well above chance. Moreover, increasing memory load made the task more difficult, so it was not too easy. This pattern therefore provides a platform to answer the main question, namely whether increasing memory load interferes with phonetic recalibration.

Memory task	% of correct probes		
	Low	Medium	High
Visuospatial	86	78	82
Verbal	98	92	83

Table 5.1. *Proportion of correctly recognized probes in the verbal and visuospatial memory task at low-, medium-, and high-memory loads.*

5.4.2 - Performance on speech identification

The data of the speech identification trials were analyzed as in previous studies by computing aftereffects (Bertelson et al., 2003; Vroomen & Baart, 2009a). First, the average number of 'b'-responses as a function of the test token was calculated for each participant. The group-averaged data are presented in Figure 5b. The data in this figure are averaged across the three memory load groups because preliminary analyses showed that memory load did not affect performance in any rational way (all F's with load as factor < 1). As is clearly visible, there were more 'b'-responses for the 'b-like' A?-1 token than the more 'd-like' A?+1 token. More interestingly, there were more 'b'-responses after exposure to A?Vb than A?Vd (indicative of recalibration), whereas there were fewer b-responses after exposure to AbVb than AdVd (indicative of selective speech adaptation), thus replicating the basic results for recalibration and selective speech adaptation reported before. To quantify these aftereffects, the proportion of 'b'-responses following exposure to Vd was subtracted from exposure to Vb, thereby pooling over test tokens. Recalibration (A?Vb – A?Vd) manifested itself as more 'b'-responses following exposure to A?Vb than A?Vd, whereas for selective speech adaptation (AbVb – AdVd), there were fewer 'b'- responses after exposure to AbVb than AdVd (see Table 5.2). Most importantly, none of these aftereffects was modulated by either of the two secondary memory tasks. This was tested in a 2 (adapter sound: ambiguous/non-ambiguous) x 3 (task: no/visuospatial/verbal) x 3 (memory load: low/medium/high) ANOVA on the aftereffects with memory load as a between-subjects

variable, and adapter sound and task as within-subjects variables. There was a main effect of adapter sound because exposure to the ambiguous adapter sounds induced positive aftereffects (recalibration), whereas exposure to the non-ambiguous sounds induced negative aftereffects (selective speech adaptation), $F(1,64) = 27.33$, $p < .001$. Crucially, there was no effect of task; $F(2,128) < 1$, memory load; $F(1,64) < 1$, nor was there a higher order interaction between any of these variables (all p's were at least > .3). Aftereffects indicative of recalibration and selective speech adaption were thus *un*affected by whether participants were trying to remember letters or a visuospatial path during the exposure and test phase.

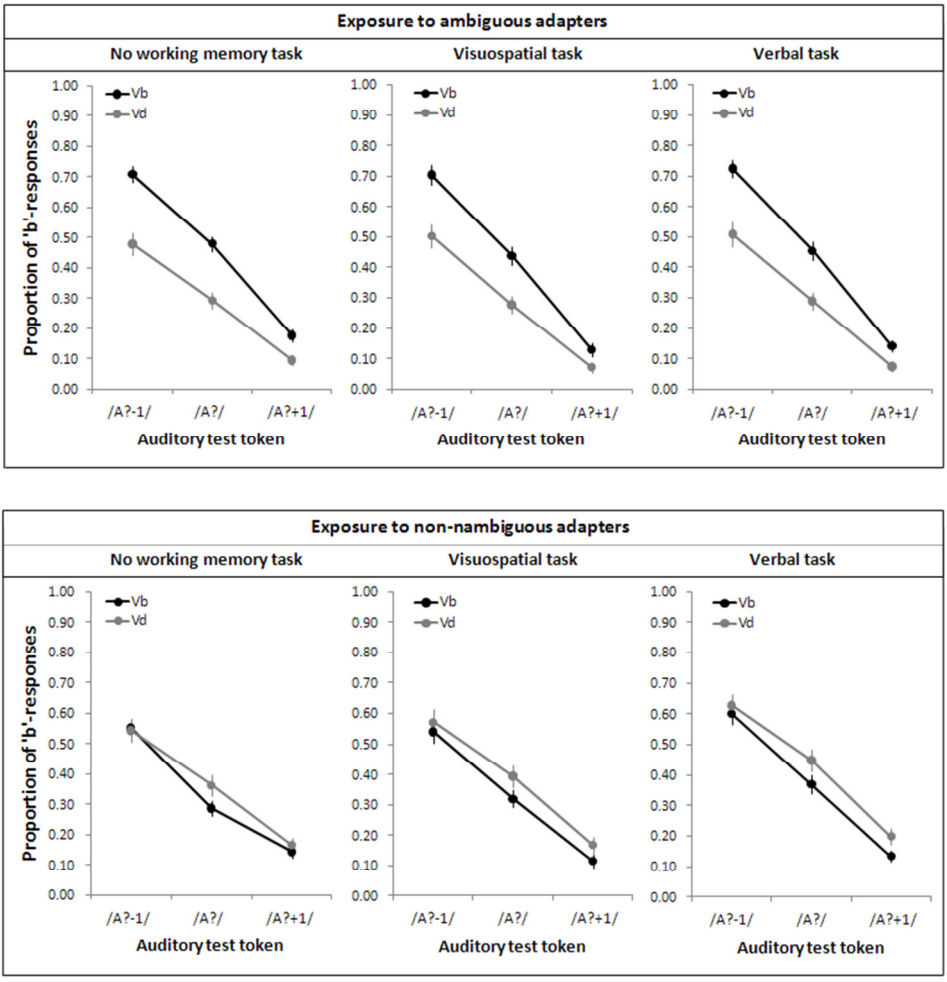

Figure 5b. *Proportion of 'b'-responses after exposure to A?Vb and A?Vd (upper panels) and AbVb and AdVd (lower panels) for the single and dual tasks. Data are averaged over memory load. Error bars represent one standard error of the mean.*

	Ambiguous adapter sound			Non-ambiguous adapter sound		
		Load			Load	
Memory task	Low	Medium	High	Low	Medium	High
No task	.15	.18	.16	-.04	-.04	-.02
Visuospatial	.15	.14	.12	-.08	-.05	-.02
Verbal	.14	.11	.17	-.07	-.06	-.05

Table 5.2. *Aftereffects after exposure to ambiguous and non-ambiguous adapter sounds while remembering verbal or spatial items at three loads.*

5.5 - Discussion

The present study indicates that a concurrent working memory task does not interfere with lipread-induced phonetic recalibration. Participants readily adapted their interpretation of an initially ambiguous sound based on lipread information, but this occurred independent of whether they were engaged in a demanding verbal or spatial working memory task. This suggests that phonetic recalibration is - like selective speech adaptation (Samuel & Kat, 1998) - a low-level process that occurs in an automatic fashion. This finding is in line with other research that demonstrates that the on-line integration of auditory and visual speech is automatic (Besle et al., 2004; Calvert & Campbell, 2003; Campbell et al., 2001; Colin et al., 2002; Massaro, 1987; McGurk & MacDonald, 1976; Möttönen et al., 2002; Näätänen, 2001; Soto-Faraco et al., 2004).

As a counterargument, it might be argued that the memory tasks were simply too easy to affect phonetic recalibration and selective speech adaptation. Against this interpretation, though, is that increasing the memory load of the concurrent task *did* affect probe recognition. In the highest load conditions of the spatial and verbal memory task, recognition rate was at ~82%, which is well above chance level, but far from being perfect. Participants were thus likely engaged in the memory task, yet it had no effect on phonetic recalibration or selective speech adaptation.

Yet another counterargument is that one cannot be sure that participants were actively engaged in covertly repeating the memory items while they were exposed to the audiovisual speech tokens that supposedly drive recalibration. Admittedly, the critical part of the exposure phase that induces recalibration - the part in which a participant hears an ambiguous segment while seeing another phonetic segment - is very short, and there is no guarantee that participants were - at that specific time - actually engaged in

repeating the memory items. Unfortunately, we cannot offer an obvious solution for this because it is a very general problem in dual-task paradigms where there is always uncertainty about strategic effects in performing the primary and secondary task. One might, as an alternative, have used a more demanding on-line task that allows one to keep track of performance during the exposure phase. Participants might for example track a concurrent visual stimulus while being exposed to the lipread information, as this is relatively easy to measure (see e.g., Alsius, Navarra, Campbell, & Soto-Faraco, 2005). However, a disadvantage of this method is that the visual tracking task as such may interfere with lipreading, so there is interference at the sensory level rather than at the level at which phonetic recalibration occurs. Participants might thus simply not see the critical lipread information when simultaneously engaged in a visual tracking task. Other studies on audiovisual speech using this dual task have indeed found that an additional visual task (tracking a moving leaf over a speaking face) can interfere with lipreading (e.g., Tiippana, Andersen, & Sams, 2004), thus preventing any firm conclusion about whether attention affects cross-modal information integration rather than lipreading itself. A recent report on spatial attention (i.e., attending one out of two faces presented on the left and right of fixation) also confirms that endogenous attention affects lipreading rather than multisensory integration (Andersen, Tiippana, Laarni, Kojo, & Sams, 2009).

Alternatively, one could also use a secondary task that does not interfere with the auditory and visual sensory requirements of the primary task, like for instance, a tactile task. In a study by Alsius, Navarra and Soto-Faraco (2007), it was indeed reported that the percentage of illusory McGurk-responses decreased when participants were concurrently performing a difficult tactile task (deciding whether two taps were finger-symmetrical with the preceding trial). As already argued, this result by itself does not unequivocally imply that the tactile secondary task had an effect on audiovisual integration per se, because the task may also interfere with unimodal processing of the lipread information, thus before audiovisual integration did take place. However, Alsius and co-workers (2005, 2007), included auditory-only and visual-only baseline conditions in which participants repeated the word they had just heard or lipread. The authors did not find a difference in the unimodal baseline conditions between the single and dual tasks, which made them refute the idea that the secondary task affected lipreading rather than audiovisual integration. Here, we acknowledge that it remains for future research to examine whether a concurrent tactile task would also affect lipread-induced phonetic recalibration.

From a broader perspective, there is a current debate in the literature about the extent to which intersensory integration requires attentional resources. Some have

argued that intersensory integration depends on attentional resources (Alsius et al., 2005; Fairhall & Macaluso, 2009; Talsma, Doty, & Woldorff, 2007), while others have argued it does not (Bertelson, Vroomen, de Gelder, & Driver, 2000; Massaro, 1987; Soto-Faraco et al., 2004; Vroomen, Bertelson, & de Gelder, 2001; Vroomen, Driver, & de Gelder, 2001). Admittedly, the current experiment did not measure the role of attention as such, but being simultaneously engaged in two tasks is usually taken to imply that available attentional resources were divided across the two tasks. Given that there was no effect of the secondary task on lipread-induced recalibration, it appears that the present findings fit better within the perspective that multisensory integration is unconstrained by attentional resources. This finding also fits well with the observation that a face displaying an emotion has profound effects on auditory emotion-labeling but yet again, this effect occurs independent of whether or not listeners were instructed to add numbers, count the occurrence of a target digit in a rapid serial visual presentation or were asked to judge the pitch of a tone as high or low (Vroomen et al., 2001b). Similarly, in the spatial domain it has been demonstrated that vision can bias sound localization (i.e., the ventriloquist effect, e.g., Radeau & Bertelson, 1974; Bertelson, 1999), but this cross-modal bias occurs irrespective of where endogenous (Bertelson et al. 2000) or exogenous spatial attention is directed (Vroomen et al. 2001a).

To conclude, the data demonstrate that during lipread induced phonetic recalibration, the auditory and visual signals were integrated into a fused percept that left longer-lasting traces. Apparently, listeners learned to interpret an initially ambiguous sound because there was lipread information that was used to disambiguate that sound. This phenomenon is - like selective speech adaptation - likely a low-level phenomenon that does not seem to depend on processes used in spatial or verbal working memory tasks. We acknowledge, though, that at this point, the dual-task method leaves more than one interpretation open, and it appears that there is no other solution than running more experiments with different tasks.

Chapter 6

Phonetic recalibration in dyslexia [6]

[6]Adapted from:
Baart, M., de Boer-Schellekens, L., & Vroomen, J. (in prep.). Lipread-induced phonetic recalibration in dyslexia.

6.1 - Abstract

Auditory phoneme categories are less well-defined in developmental dyslexic readers than in fluent readers. Here, we examined whether poor recalibration of phonetic boundaries might be associated with this deficit. Adult dyslexic readers were compared with fluent readers on a phoneme identification task and a task that measured phonetic recalibration by lipread speech (Bertelson et al., 2003). In line with previous reports, we found that dyslexics were less categorical in the labeling of the speech sounds. The size of their phonetic recalibration effect, though, was comparable to that of normal readers. This result indicates that phonetic recalibration is unaffected in dyslexic readers, and that it is unlikely to be a cause of their impairment in auditory phoneme categorization.

6.2 - Introduction

Developmental dyslexia (henceforth DD) is characterized by substantial reading problems that cannot be explained by education, motivation, and intelligence (American Psychiatric Association, 2000). Besides their reading problems, individuals with DD often also have deficits in auditory-phonological perception, phoneme representation, and phonological memory (see e.g., Vellutino, Fletcher, Snowling, & Scanlon, 2004 for a review). Indeed, numerous studies have reported that minimally contrasting speech categories (e.g., /b/ and /d/) are less well-defined in dyslexic than fluent readers (Blomert & Mitterer, 2004; Bogliotti et al., 2008; de Gelder & Vroomen, 1998; Godfrey et al., 1981; Vandermosten et al., 2010; Werker & Tees, 1987). Human speech, though, is not only perceived by sound but also by the visual information about the articulatory movements of the mouth and face, here referred to as 'lipreading'. It has been known for a long time that in daily life, lipread information helps to improve the eligibility of auditory speech (e.g., Sumby & Pollack, 1954). It is however less well-known that lipread speech not only disambiguates ongoing auditory speech, but also has a longer-term effect on sound identification as it can 'recalibrate' existing phonetic categories. On this view, lipread information is used to 're-align' existing sound categories so that the natural correspondence between what is heard and seen is maintained.

Phonetic recalibration by lipread speech has been demonstrated in a paradigm where repeated exposure to an auditory ambiguous speech sound (i.e., from the middle of an /aba/-/ada/ continuum) in combination with clear lipread speech (i.e., a video of a speaker pronouncing either /aba/ or /ada/) elicits a shift of the phoneme boundary as measured in auditory-only post-tests (e.g., Bertelson et al., 2003). Results show that an auditory ambiguous sound halfway between /b/ and /d/ is more likely perceived as /b/ when during the previous exposure phase, the same sound was combined with lipread /b/ rather than with lipread /d/. This finding has been taken as a demonstration that listeners flexibly adjust their phoneme boundary to include an ambiguous sound into a particular speech category based on previously encountered lipread information (e.g., Baart & Vroomen, 2010b; Bertelson et al., 2003; van Linden & Vroomen, 2007; Vroomen et al., 2007; Vroomen et al., 2004).

Given that previous studies have reported that individuals with DD may be impaired in phonetic sound categorization, we though it important to examine to which extent DD-related deficits in auditory phoneme categorization are associated with poor phonetic recalibration. Of course, one can ask why one would expect a link between sound categorization and recalibration in the first place. We would argue that, in general, the phonetic speech recognition system has to deal with two quite different

requirements: On the one hand, it needs to be *precise* to make fine-grained distinctions between sounds that can be very similar, like the difference between a /b/ and a /d/. On the other hand, it also needs to be *flexible* so that it can adjust to acoustic variations between different utterances, speakers, environments, and so forth. Traditionally, these two requirements (precision and variability) have been studied in isolation (see for instance Samuel, 2011, for a review), but it seems plausible that both need to be handled at the same time while speech sounds are processed. It seems logical then, that a well-calibrated system can make better distinctions than a poorly calibrated system. For this reason, we expected listeners to be more precise in sound categorization the better they were able to calibrate their phonetic system.

Starting from the observation that dyslexic readers have poor sound categorization, one can envisage various links between this skill and phonetic recalibration by lipread speech. One possibility, already alluded to, is that poor sound categorization emanates from a deficit in the ability to flexibly adapt the system. One cause of poor recalibration by lipread speech might be that lipreading itself is compromised. Some have indeed reported that poor readers are also poor lipreaders (e.g., de Gelder & Vroomen, 1998; Mohammed et al., 2006). For instance, Mohammed et al. (2006) showed a deficit in lipreading in adult dyslexic readers when asked to match a lipread word, sentence, or short story with a picture. Compromised lipreading skills would then hinder necessary cross-modal adjustments of auditory speech input, resulting in poorly-defined auditory speech representations. Alternatively though, auditory impairments might equally-likely be the primary cause of any deficiencies in sound categorization, as DD-related auditory speech deficits have been shown to already be present at birth. For example, newborns with familial risk for dyslexia display deviant brain activity if compared to non-risk infants when presented with synthetic /ba/, /da/ and /ga/ sounds (Guttorm et al., 2003; Leppanen et al., 1999), which in turn is closely related to poor receptive language skills and verbal memory in the following years of development (Guttorm et al., 2005). Lipreading skills are also known to develop with age (e.g., Massaro, 1984; McGurk & MacDonald, 1976) and another possibility is that this developmental trend as such is disrupted by the DD-related auditory impairments. Yet another possible link is that speech-specific perceptual problems in dyslexia are restricted to auditory-only speech as dyslexics may have learned to compensate for their auditory deficits by relying more on lipread input. This idea is in line with a report showing that dyslexics display enhanced brain activity in areas dedicated to visual- and motor-articulatory processes as compared to controls when presented with audiovisual speech (Pekkola et al., 2006).

Here, we did not have the ambition to resolve all these issues. Rather, as a first approximation, we tried to obtain data on whether dyslexic readers actually calibrate their phonetic system like fluent readers do and additionally, we sought to obtain empirical evidence for the suggestion that there is a relation between sound categorization and recalibration. For this, we adopted a paradigm described in Bertelson et al. (2003). Listeners were repeatedly exposed to a short block of audiovisual adapters that contained an auditory ambiguous sound halfway between /b/ and /d/ (the sound closest to the individually determined phoneme boundary, henceforth A?) that was combined with lipread (visual) information of /b/ or /d/, thus yielding A?Vb and A?Vd, respectively. After a short exposure block to A?Vb or A?Vd, participants were tested on their identification of auditory-only sounds near their phoneme boundary. Recalibration should manifest itself as a higher likelihood to label the ambiguous sound as /b/ after exposure to A?Vb than after exposure to A?Vd. As a control for simple perseveration or priming effects, we included, as in previous studies, auditory non-ambiguous and audiovisual congruent exposure stimuli AbVb and AdVd. These stimuli typically yield no recalibration effect – because there is no conflict between what is heard and seen – but may yield a relatively small contrastive aftereffect due to selective speech adaptation because of the non-ambiguous nature of the sound (e.g., Eimas & Corbit, 1973). We expected that the DD-group would be less categorical than the controls in identifying auditory-only sounds, reflecting poorer-defined /b/-/d/ speech categories. The critical question was whether the dyslexics would also display different recalibration effects, and whether there was correlation between these two measures.

6.3 - Method

6.3.1 - Participants

12 students (8 females) from Tilburg University, formally assessed and diagnosed with DD and 12 gender- and age-matched controls (also students) participated. All participants were native Dutch speakers between 18 and 25 years of age (Mean age was 21 years in both groups, $t(22) = .093$, $p = .927$). All reported normal hearing, had normal/corrected to normal vision, and gave their written informed consent prior to testing. All testing was conducted in accordance with the Declaration of Helsinki.

Non-verbal IQ (Raven's Standard Progressive Matrices test, Raven, Raven, & Court, 1998) could be assessed in half of the participants from each group, and showed no group difference (raw score was 50.00 for the dyslexic group vs. 49.50 for the controls, $t(10) = .182$, $p = .860$). Before testing started, all participants were given two Dutch standardized tests that measured single word reading for real words and pseudo-

words, namely the 'Een-minuut-test' (i.e., EMT, Brus & Voeten, 1997) and 'De Klepel' (Van den Bos, Lutje Spelberg, Scheepsma, & De Vries, 1999). As expected, reading scores were lower for the DD- than the control group (70.4 versus 99.0, $t(22) = 5.33$, $p < .001$ for words, and 66.3 versus 100.3, $t(22) = 6.50$, $p < .001$ for pseudo-words, respectively).

6.3.2 - Stimuli

Stimulus details are described in an earlier paper (Bertelson et al., 2003). In short, audiovisual recordings of a male speaker of Dutch pronouncing the pseudo words /aba/ and /ada/ were made (at 25 frames/second). The audio was synthesized into a nine-token /aba/ - /ada/ continuum by changing the second (F2) and third (F3) formants in eight equal Mel-steps using the 'Praat' speech editor (Boersma & Weenink, 2005). To ensure accurate timing between sound and vision, videos were displayed as two strings of bitmaps (each bitmap displayed for 40 ms at a refresh rate of 100 Hz) while the sound was delivered by trigger, thus preserving the original timing. To induce recalibration, the individually determined most ambiguous sound of the continuum (A?) was combined with the video of /aba/ (Vb) or /ada/ (Vd), resulting in two audiovisual adapters; A?Vb and A?Vd. As a control, we included the audiovisual congruent adapters AbVb and AdVd that consisted of the auditory non-ambiguous endpoints of the continuum with the corresponding video.

6.3.3 - Design and procedure

Participants were tested individually in a dimly lit and sound attenuated booth. Participants sat at approximately 70 cm from a 17-inch CRT-monitor. The audio was delivered at ~62 dBa (ear level) via two regular computer speakers (JBL Media 100).

The total experimental procedure lasted about 45 minutes and consisted of four phases: a silent visual /b/-/d/ discrimination task, an auditory /b/-/d/ identification task to test sound categorization and determine the individual phoneme boundary, an exposure – test phase to test recalibration, and an auditory goodness rating task of the audiovisual adapters.

6.3.3.1 - Silent visual speech discrimination

To test whether there was any difference in discriminating lipread /aba/ from /ada/, we delivered Vb and Vd 12 times each in random order without sound. Following each stimulus presentation, participants decided whether they saw /aba/ or /ada/ being pronounced by pressing the corresponding key on a keyboard. The next stimulus was delivered 750 msec after key-press.

6.3.3.2 - Auditory identification

The nine auditory tokens of the continuum were presented 12 times each in random order. On each trial, participants watched a fixation cross on the screen and indicated whether they heard /aba/ or /ada/ by pressing the 'b'- or 'd'-key. The next trial started 1000 msec after detection of the key-press. After testing, the perceptually most ambiguous token of the continuum was determined for each participant. This was done by fitting a cumulative function on the proportion of 'b'-responses. The stimulus closest to the 50% cross-over point served as the most ambiguous token (A?) in the following recalibration phase.

6.3.3.3 - Exposure-test phase

Participants were repeatedly presented a short exposure block of audiovisual adapter stimuli followed by six auditory-only test trials. Each exposure block consisted of eight repetitions (ISI = 150 msec) of one of the four audiovisual adapters A?Vb, A?Vd, AbVb, and AdVd. Exposure was immediately (400 msec) followed by an auditory-only test in which participants indicated whether they heard /aba/ or /ada/ by pressing a corresponding key. The auditory test stimuli were A?, its more 'aba-like' neighbour on the continuum (A?-1), and the more 'ada-like' neighbour on the continuum A?+1. These three auditory tokens were delivered twice each in random order (six test trials, ITI = 1000 msec). In total, 32 of these short exposure – test blocks were delivered in random order (8 blocks per audiovisual adapter). To ensure that participants attended the screen during exposure, occasional catch-trials consisting of a small white dot above the upper lip of the speaker (Ø ~3 mm, 120 msec in duration) had to be detected by pressing a designated key.

6.3.3.4 - Goodness rating of audiovisual adapters

To ensure that the exposure stimuli A?Vb, A?Vd, AbVb, and AdVd were perceived in a similar way, we asked participants at the end of the experiment to rate the /b/-/d/ quality of the auditory signal of the audiovisual adapters on a 7-point Likert-scale with '1' representing a clear /b/ and '7' a clear /d/. The next trial started 1200 msec after key press. All four adapters were presented eight times in pseudorandom order (32 trials in total).

6.4 - Results

6.4.1 - Discrimination of lipread stimuli

The data of the discrimination task for lipread material were analyzed by measuring the proportion of correct responses (a 'b'-response after Vb, and a 'd'-

responses after Vd). A 2 (Video identity: Vb vs. Vd) x 2 (Group: DD's vs. controls) ANOVA on these data showed no main effect of video-identity (F-value < 1) as both videos were correctly identified in 98% of the trials. There was no main effect of group and no interaction between video-identity and group (both F-values < 1), thus indicating that discrimination of lipread /b/ from /d/ was alike in both groups and at ceiling.

6.4.2 - Auditory identification

For the auditory identification test, we measured the proportion of 'b'-responses for each token of the continuum. A 9 (Auditory token) x 2 (Group) ANOVA showed a main effect of auditory token ($F(8,176) = 255.04$, $p < .001$) because unsurprisingly, there were more 'b'-responses for the more 'b-like' tokens of the continuum. The overall proportion of 'b'-responses was also lower for the DD-group than for the controls (.42 vs. .51, $F(1,22) = 6.28$, $p < .020$), and there was an interaction between the auditory token and group ($F(8,176) = 2.77$, $p < .007$. To examine this in more detail, we fitted a logistic function on the individual raw data (see Figure 6a).

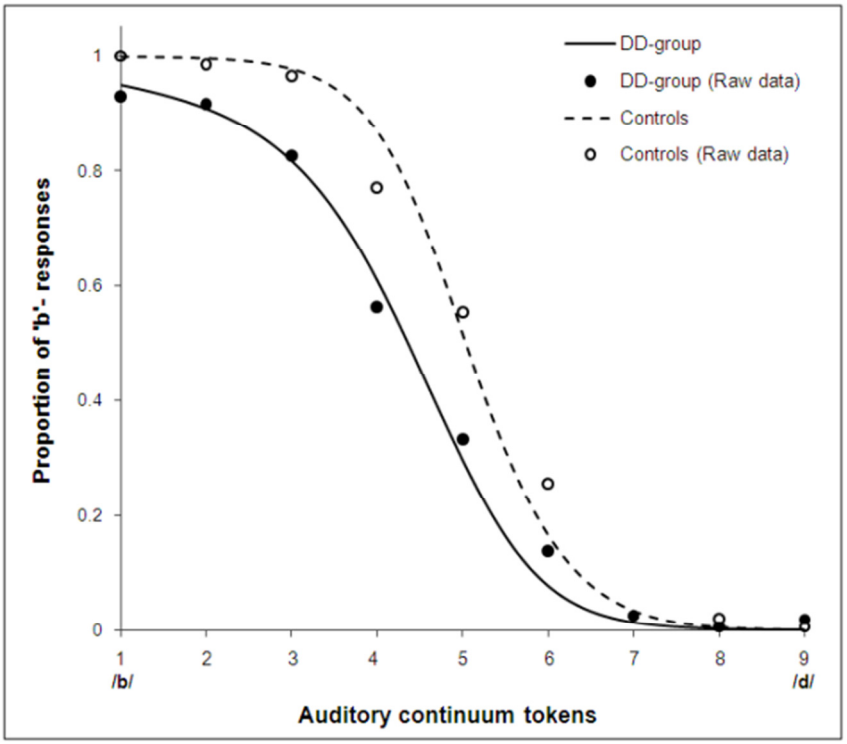

Figure 6a. *Proportion of 'b'-responses on the auditory continuum tokens for the DD- and control group.*

This allowed us to determine the 50% cross-over point, reflecting the /b/-/d/ phoneme boundary, and the just noticeable difference (JND). The JND is an indication of the smallest sound-interval that participants can reliably notice (> 75%), and can be regarded as a measure of 'categoricalness' in phoneme identification. The analyses showed that the DD-group had their phoneme boundary located more towards the /b/-end of the continuum (at 4.13 stimulus units) than the control group (5.07 units), $t(22) = 2.61$, $p < .016$. As expected, the average JND in the DD-group was larger (.68) than in the control group (.57), ($t(22) = 1.93$, $p_{one\text{-}tailed} < .034$, thus indicating that the DD-group was less categorical in labelling the continuum sounds than the controls.

6.4.3 - Exposure - test

The critical data of the exposure - test phase are presented in Table 6.1. The data were analyzed as in previous studies by computing aftereffects (e.g., Bertelson et al.,2003), thereby pooling the proportion of 'b'-responses over the three test tokens (see Table 6.1). As expected, after exposure to auditory ambiguous sounds there were substantially more 'b'-responses after exposure to A?Vb than A?Vd, reflecting the recalibration effect (a 35% overall difference). For the auditory non-ambiguous control adapters AbVd and AdVd, there was no difference and listeners were equally likely to report /b/ after exposure to AbVb or AdVd. Most importantly, these aftereffects did not differ between the two groups.

		Visual information		
Group	Sound quality	Vb	Vd	Aftereffect
DD	Ambiguous	.55	.22	.33*
	Non-ambiguous	.44	.47	-.03
Control	Ambiguous	.57	.21	.36*
	Non-ambiguous	.46	.47	-.01

*$p < .002$

Table 6.1. *Proportion of 'b'-responses for the DD-group and the controls after exposure to four different audiovisual adapters and the corresponding aftereffect.*

These generalizations were confirmed in a 2 (Sound ambiguity: auditory ambiguous vs. auditory non-ambiguous adapters) x 2 (Group) ANOVA on the aftereffects. There was a main effect of sound ambiguity ($F(1,22) = 70.62$, $p < .001$), with no main effect of group, and no interaction between the two factors (F-values < 1). Separate t-test confirmed that aftereffects were bigger than zero for adapters containing ambiguous sounds, reflecting recalibration (DD: a 33% aftereffect, $t(11) = 9.35$, $p < .001$; Controls: a 36% aftereffect, $t(11) = 4.28$, $p < .002$), but not for adapters containing auditory non-ambiguous sounds (both p-values > .4).

We also examined whether there was a correlation between the size of the lipread-induced recalibration effect and the categoricalness in labeling the continuum sounds. For normal readers, recalibration effects were negatively correlated with their auditory JND ($r = -.60$, $p < .040$), thus demonstrating that more categorical perceivers (a small JND) had larger recalibration effects. For the DD-group, this correlation was not significant ($r = -.12$, $p = .70$). Due to the relatively small group sizes, though, the difference between the two correlations (after a Fisher z transformation) did not reach significance ($z = 1.20$, $p = .115$).

6.4.4 - Goodness ratings

To analyze the auditory goodness ratings of the exposure stimuli, we computed the average rating per adapter (see Table 6.2). The 2 (Sound ambiguity) x 2 (Video identity) x 2 (Group) ANOVA showed no main effect of sound ambiguity (F < 1). As expected, there was an effect of video identity ($F(1,22) = 671.78$, $p < .001$) because videos containing /aba/ were rated more 'b'-like than videos containing /ada/ (1.46 vs.

		Visual information		
Group	Sound quality	Vb	Vd	Difference
DD	Ambiguous	1.70	5.93	4.23
	Non-ambiguous	1.21	6.68	5.47
Control	Ambiguous	1.85	5.73	3.88
	Non-ambiguous	1.10	6.47	5.37

Table 6.2. *Goodness ratings on a 7-point Likert-scale of the four audiovisual adapters and the obtained difference scores for the DD-Group and the controls.*

6.20 respectively). This effect was modulated by adapter ambiguity ($F(1,22) = 31.66$, $p < .001$) as the auditory unambiguous adapters were rated more towards the endpoints of the scale (i.e., as better examples) than the adapters containing auditory ambiguous sounds. The ANOVA showed no main effect of group, nor did group interact with any (combination) of the other factors (all F-values < 1). This result indicates that the audiovisual adapters were perceived in a similar fashion by dyslexic and normal readers.

6.5 - Discussion

Dyslexic readers were compared with fluent readers on a /b/-/d/ sound identification task and a task that measures phonetic recalibration by lipread speech. The data regarding sound identification demonstrated that dyslexic readers were less categorical in labeling the speech sounds from the /b/-/d/ continuum than the control group. This result confirms previous studies that indicate that dyslexic readers have poorer-defined phonetic sound categories than fluent readers (e.g., Bogliotti et al., 2008; de Gelder & Vroomen, 1998; Godfrey et al., 1981; Vandermosten et al., 2010; Werker & Tees, 1987). The new finding here is that phonetic recalibration by lipread information was, in essence, intact in the DD-group as the amount of recalibration was comparable in size with that of normal readers. At first sight, it thus seems conceivable that the dyslexics' deficits in the categorization of auditory speech are unlikely to originate from an inability to recalibrate the phonetic system.

There are, however, several caveats that need to be taken into account. First, besides the usual problems with interpreting a null-effect, it should be realized that we tested only a relatively small group of university students who are unlikely to be representative for all dyslexic readers. It is also interesting to note that in the normal readers, but not in the DD-group, there was a link between the size of the recalibration effect and the categoricalness of the labeling of the speech sounds. We argued that there are theoretical reasons why categoricalness in sound identification might be linked with phonetic recalibration, viz. sound identification can be more sensitive the better the phonetic system is calibrated. In normal readers, we indeed found this correlation, as individuals with well-defined speech categories (i.e., a small JNDs) had large lipread-induced aftereffects (i.e, a large recalibration effect). For the DD group, though, this correlation was not significant. It remains therefore necessary to test more subjects before any firm conclusions can be drawn that dyslexics have, in general, normal phonetic recalibration.

Another interesting finding was that the DD-group was not impaired in the visual-only discrimination of lipread /b/ from /d/, as the visual-only performance of both

groups was alike and almost flawless (98% correct). The goodness ratings of the audiovisual adapters also showed that both groups were equally affected by the visual input. Most likely then, both groups were equally good in lipreading the stimuli used here. This may seem remarkable because in a previous study on the recognition of audiovisual speech, de Gelder and Vroomen (1998) actually used the same phonetic /b/-/d/ contrast (although different stimuli) and reported considerably lower proportions of correctly lipread responses in both a poor- (.67) and a normal reading group (.77). A potentially relevant difference though, is that this study tested dyslexic children rather than adults. It is well-known that children are less proficient in decoding lipread speech (e.g., McGurk & MacDonald, 1976). Moreover, this developmental trend in the effective use of lipread information is further underscored by a more recent study that used the same stimuli and procedures as here and showed that lipread-induced phonetic recalibration develops with age (van Linden & Vroomen, 2008). It seems therefore conceivable that lipread-induced recalibration of auditory speech is related to the extent that perceivers are actually able to lipread the stimuli.

To conclude, the combination of normal recalibration with compromised auditory categorization suggests that dyslexia-related impairments in auditory phoneme categorization, most likely, do not originate from an inability to recalibrate the phonetic system. However, it remains for future studies to explore whether this is also the case in a wider sample of dyslexics. It could be argued that students show a compensated dyslexic profile with milder literacy and language deficits than those typically observed in a larger dyslexic population. One possible way to tap into the critical processes would be by testing a group of dyslexic children rather than adults. One caveat, though, is that one needs to take into account that there is a developmental trend in the use of lipread information that might easily confound lipread-induced recalibration effects obtained with children.

Chapter 7

Audiovisual phonetic binding [7]

[7]Adapted from:

Baart, M., Stekelenburg, J. J., & Vroomen, J. (in prep.). Perception of Audiovisual Sine-wave speech: ERP evidence for a Phonetic Mechanism at P2.

7.1 - Abstract

EEG studies have shown early visually induced modulations of the auditory N1 response that have been interpreted as evidence for early speech specific AV integration processes. We used audiovisual sine-wave speech stimuli that were only perceived as speech by half of the participants and observed similar modulations of N1 for all participants supporting the alternative view that the auditory N1 is sensitive to visual anticipatory motion and not restricted to speech processing. We additionally observed lipread induced P2 modulations, but only for listeners that perceived the sounds as speech. Later effects of stimulus congruency were also obtained, but again, were restricted to the speech group. We suggest that the time-course of AV speech perception reflects at least three subsequent levels of integration in which the natural temporal characteristics are integrated (at N1) before the modality inputs are bound phonetically (at P2) and stimulus congruency is processed and perceptually finalized.

7.2 - Introduction

The processing of an auditory speech signal is known to be influenced by the corresponding visual speech input of the articulatory gestures of the talker (here referred to as 'lipreading'). For instance, seeing the video of a talking face helps to correctly identify a speech sound masked in noise (Sumby & Pollack, 1954). Additionally, an auditory /ba/ sound combined with lipread /ga/ will be perceived as a /da/ because the audiovisual (hence AV) perceptual conflict is solved by fusing the inputs from both modalities into one percept (McGurk & MacDonald, 1976). By using Electroencephalography (EEG), it has been demonstrated that lipread speech affects auditory processing in the auditory cortex as early as ~100 – 200 msec after sound onset (Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005). More specifically, the auditory N1 component in the recorded Event Related Potentials (ERPs) is attenuated (Besle et al., 2004; Klucharev et al., 2003) and speeded-up (van Wassenhove et al., 2005) by simultaneously delivered lipread input.

Although these lipread-induced auditory N1 modulations were originally proposed to be specific for AV speech processing, there is accumulating evidence in support of a different view: Stekelenburg and Vroomen (2007) demonstrated that ecologically valid stimuli such as videos of an actor who is clapping hands or a spoon hitting a cup, that are combined with the corresponding sounds also elicit a visually-induced speeding-up and attenuation of N1. Given that in AV speech, lipread input usually precedes the auditory signal (e.g., Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009), Stekelenburg and Vroomen (2007) argued that the visually-induced N1 modulations arise whenever the visual input precedes the audio, thus warning the listener about when the sound is going to occur. This was corroborated by similar results obtained with artificial AV stimuli in which anticipatory visual motion reliably predicted sound onset (Vroomen & Stekelenburg, 2010). It thus appears that the temporal characteristics of an AV stimulus in which the visual component precedes the audio are responsible for the early N1 effects, irrespective of whether the stimuli are ecologically valid or artificial.

Although the visually-induced N1 modulations are not exclusive for speech, this does not imply that there are no specialized mechanisms dedicated to AV speech processing. In fact, it seems likely that AV speech integration is a multi-staged process in which separate features, including phonetic information, get integrated on different levels of processing (e.g., Eskelund, Tuomainen, & Andersen, 2011). On this view, the natural temporal characteristics of the AV speech signal may provide integration on one particular level, as for instance demonstrated by the early N1 modulations. Interestingly, converging results from Magneto encephalography (MEG) and functional Magnetic

Resonance Imaging (fMRI) have revealed a fast and direct predictive mechanism from visual to auditory brain areas (Arnal, Morillon, Kell, & Giraud, 2009).

Phonetic AV binding is however likely to be constituted on a different level as for instance demonstrated in a recent study by Eskelund et al. (2011) in which the authors used so-called Sine-wave speech (hence SWS). In SWS, the natural spectral richness of the signal is reduced to sinusoids that follow the centre frequency and the amplitude of the first three formants. Typically, naïve listeners perceive SWS as 'non-speech' sounds like whistles or computer bleeps. However, once listeners are told that these sounds are derived from speech (i.e., when they are in 'speech mode'), they cannot switch back to a non-speech mode again and continue to hear the sounds as speech (Remez et al., 1981). Eskelund et al. (2011) reported a lipread-induced detection benefit for the auditory stimulus for all participants, irrespective of whether they were in speech- or non-speech mode. Critically, phonetic integration was only found when listeners were in speech mode, in close correspondence with earlier reports (Tuomainen et al., 2005; Vroomen & Baart, 2009a; Vroomen & Stekelenburg, 2011). fMRI studies have revealed that activity in the Superior Temporal Sulcus (STS) increases when SWS is perceived as speech (Benson, Richardson, Whalen, & Lai, 2006; Möttönen et al., 2006) and that during audiovisual integration of SWS stimuli, sensitivity to stimulus intelligibility and linguistic access are reflected in anterior regions of STS (Lee & Noppeney, 2011).

The advantage of SWS-stimuli is that they allow a distinction between speech and non-speech processing while keeping visual predictive information and acoustic properties the same for all listeners. SWS thus provides an ideal platform to investigate whether the underlying time-course of speech processing is different from non-speech processing. Here, this hypothesis was tested in an EEG paradigm where we recorded ERPs while presenting SWS stimuli in auditory-only, visual only (i.e., silent videos of a speaker) and in AV fashion in two groups of listeners; a speech- and a non-speech group. Assuming that visual anticipatory information is critical for the N1 modulations, we expected that visual speech would induce an attenuation and speeding-up of the N1 in both groups.

We additionally did expect a difference, though not at N1, in the lipread-induced modulations of the auditory ERPs for the speech and non-speech groups and the time-frame at which this difference occurs might be indicative for the time at which lipread speech can modulate phonetic processing.

An ERP component that could potentially reflect this speech-specific mechanism is the positive peak at ~200 msec (i.e., the P2). Previous work has demonstrated that, in addition to the N1 modulations, lipread speech elicits an

attenuation of the auditory P2 (Stekelenburg & Vroomen, 2007). Although the auditory P2 reduction by visual input can also be elicited with artificial stimuli, it is argued to be functionally dissociated from N1 (Vroomen & Stekelenburg, 2010).

Interestingly, the P2 appears to be sensitive to a violation of expected temporal, semantic and/or phonetic AV congruency (Stekelenburg & Vroomen, 2007). However, the exact properties of the stimuli that elicit P2 modulations are currently unknown. Here, we sought to tease apart P2 modulations that are potentially triggered by a phonetic mechanism from modulations induced by the detection of incongruency by using stimuli in which the AV incongruency occurred too late (> 270 msec) to elicit P2 modulations. We hypothesized that, if a phonetic mechanism is reflected at P2, only listeners in the speech group would show lipread induced P2 modulations as the non-speech group makes no use of the phonetic content of the lipread signal (Tuomainen et al., 2005; Vroomen & Baart, 2009a; Vroomen & Stekelenburg, 2011). If so, this would suggest that the earliest phonetic mechanisms specific to AV speech are reflected in the ERPs at around 200 msec rather than at ~100 msec as previously argued (Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005).

7.3 - Method

7.3.1 - Participants

28 first-year students (20 females) from Tilburg University participated in return for course credits. Half of them were randomly assigned to the 'speech' group, the other half to the 'non-speech' group. Participants' age ranged in between 18 and 26 years (Mean = 21) and did not differ across groups ($t(26) = 1.17$, $p = .252$). All reported normal hearing, had normal/corrected to normal vision, and gave their written informed consent prior to testing. All testing was conducted in accordance with the Declaration of Helsinki.

7.3.2 - Stimuli

Stimulus material is described in detail in an earlier paper (Vroomen & Stekelenburg, 2011). In short, audiovisual recordings of a Dutch male speaker pronouncing the pseudo-words /tabi/ and /tagi/ were made (25 frames/s). The audio was converted into sine-wave speech via a script provided by C. Darwin (http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/SWS) in the Praat software (Boersma & Weenink, 2005). Videos displayed the speakers' face from shoulders up to crown. The audio files were 627 msec (/tabi/) and 705 (/tagi/) msec in duration and onsets of the critical consonants were at 270 (/b/) and 300 (/g/) msec. For experimental purpose, eight different stimuli were devised; the auditory-only SWS /tabi/

and /tagi/ sounds (i.e., Ab and Ag), both visual-only videos (Vb and Vg), two audiovisual congruent (AbVb and AgVg) and two incongruent (AbVg and AgVb) stimuli.

7.3.3 - Procedure and design

Participants were tested individually in a dimly lit and sound attenuated booth. Participants sat at approximately 70 cm from a 17-inch CRT-monitor. The audio was delivered at ~65 dBa (ear level) via a regular computer speaker placed directly below the monitor. Size of the videos subtended 14° horizontal and 12° vertical visual angle. Accurate timing between sound and vision was preserved by displaying the videos as a string of bitmaps (each bitmap displayed for 40 msec at a refresh rate of 100 Hz) while the sound was delivered by trigger. For the incongruent presentations, audiovisual stimuli looked and sounded naturally timed as the 30 msec difference in /b/ versus /g/ onsets is well within the critical AV binding window (van Wassenhove, Grant, & Poeppel, 2007).

The experiment started with a short training. Participants in speech mode learned to perceive the SWS stimuli as speech in a procedure were presentations of the original audio recordings and the corresponding SWS tokens were alternated (twelve times each) whereas the non-speech group heard only the SWS sounds (also 12 times for each sound) while under the impression they were hearing two different arbitrary computer sounds. After training, none of the participants in non-speech mode reported hearing the sounds as speech when asked to describe the sounds. Next, ERPs were recorded during six ~10-minute blocks with short breaks in between. One experimental block comprised 96 experimental trials and 16 catch trials delivered in random order. Half of the experimental trials were unimodal and the other half were audiovisual. Half of the unimodal trials were auditory-only (i.e., 12 Ab and 12 Ag trials) and the other half were visual-only trials (12 Vb and 12 Vg trials). Of the audiovisual trials, 24 were congruent (12 AbVb and 12 AgVg trials) and 24 were incongruent (12 AbVg and 12 AgVb trials). Participants were all engaged in an unrelated visual detection task: They were instructed to attend to the stimuli and press a button when an occasional small white square appeared (120 ms) on the upper-lip of the speaker (or on the black screen during auditory-only trials). Two presentations of each of the 8 different stimuli were such catch trials.

7.3.4 - EEG recording and analysis

The electroencephalogram (EEG) was recorded at a sampling rate of 512 Hz from 64 locations corresponding to the extended International 10-20 system. Electrodes

were active Ag–AgCl electrodes (BioSemi, Amsterdam, the Netherlands), mounted in an elastic cap. Two additional electrodes served as reference (Common Mode Sense active electrode; CMS) and ground (Driven Right Leg passive electrode; DRL) and two additional electrodes were placed on the left and right mastoids. EEG was referenced off-line to an average of these mastoid electrodes and band-pass filtered (0.1-30 Hz, 24 dB/octave). ERPs were time-locked to auditory onset and the raw data were segmented into epochs of 900 ms, including a 100-ms pre-stimulus baseline. After EOG correction (Gratton, Coles, & Donchin, 1983), epochs with an amplitude change > 120 μV at any EEG channel were rejected.

ERPs of the catch trials were excluded from analyses and the remainder of the trials were averaged per modality (A, V and AV) for both groups separately. In line with earlier reports, (e.g., Besle et al., 2004; Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010), this allowed us to compare the audiovisual (AV – V) with the auditory-only (A) condition and interpret any difference as an integration effect between the two modalities. In a first analysis, we compared both groups on the N1 and P2 peaks to reveal lipread-induced modulations that reflect visual prediction (N1) and possibly, a phonetic mechanism (P2). Since the auditory N1 and P2 have a central topography, analyses were performed on the central electrode Cz. The peak of the N1 was determined within in a window of 70–150 msec, and the P2 peak was scored in a window of 150–250 msec.

Assuming that only listeners in the speech group would integrate the auditory and visual information at a phonetic level (Eskelund et al., 2011; Tuomainen et al., 2005; Vroomen & Baart, 2009a; Vroomen & Stekelenburg, 2011), we anticipated that the effect of AV-congruency might be different across groups. To investigate this hypothesis, we constructed difference waves by subtracting the AV congruent ERPs from the incongruent ones and made between group comparisons as specified below.

7.4 - Results

Participants were almost flawless on catch trial detection (99% in the speech group versus 100% in the non-speech group, $t(26) = 1.13$, $p = .268$) indicating that they were indeed looking at the screen as instructed. The ERP data of the speech group and the non-speech group are shown in Figure 7a. As expected, sounds induced a clearly visible auditory N1 whose amplitude was maximal at Cz. Figure 7a suggests that adding lipread information in the AV-condition sped up and reduced the amplitude of the N1 in both groups alike. The amplitude of the P2 was also reduced by lipread information, and this effect was bigger for the speech than the non-speech group.

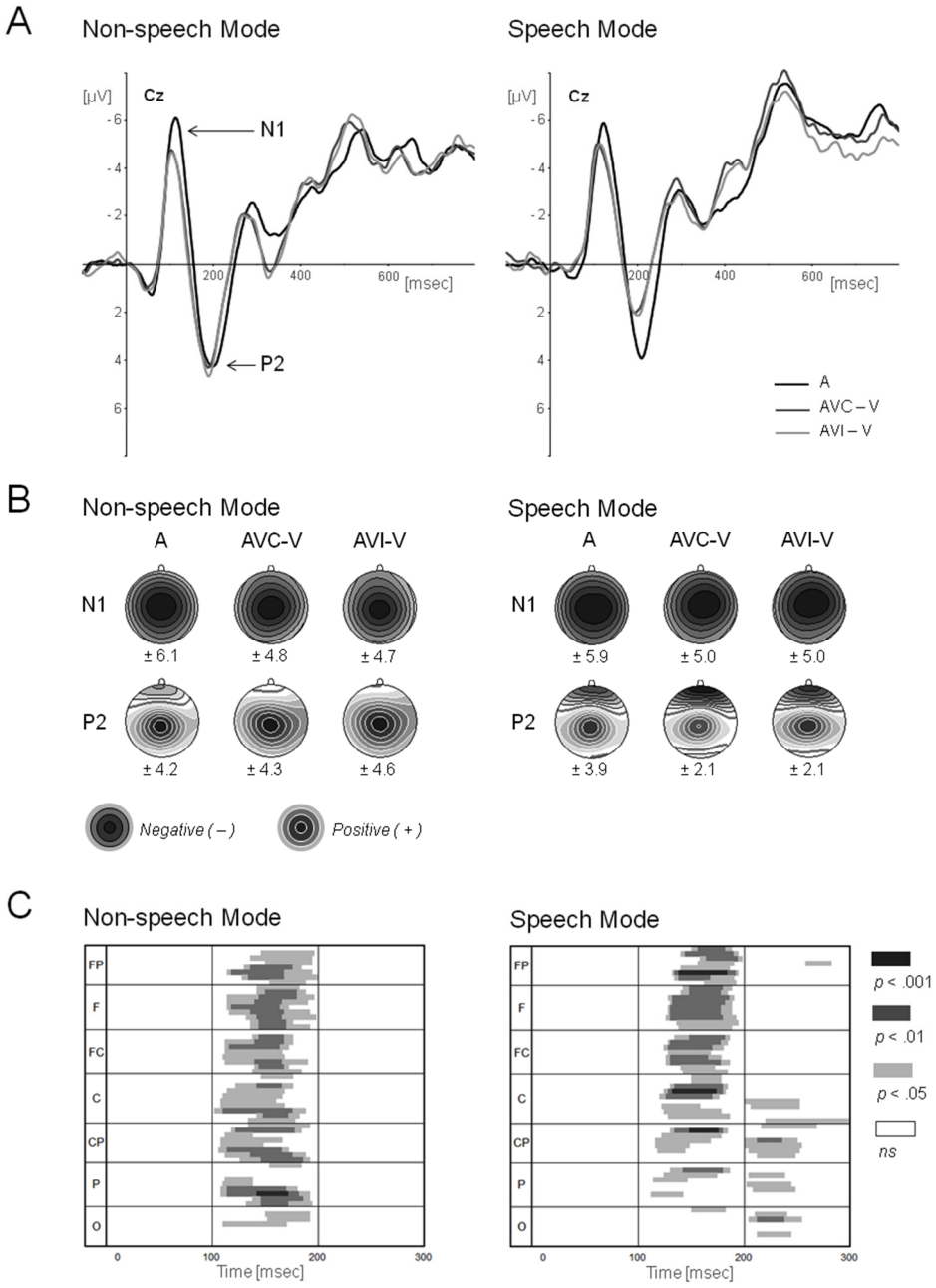

Figure 7a. *Event-related potentials (ERPs) at electrode Cz (A) and the scalp topography including the (range of the) voltage maps of the N1- and P2 peaks (B) for the auditory-only, the AV congruent (AVC – V) and AV incongruent (AVI – V) condition. ERPs for the speech- and non-speech group were pooled across /tabi/ and /tagi/. Figure 7aC displays the time course of the AV interactions of the congruent stimuli (AV – V – A) using point wise t-tests at every electrode.*

7.4.1 - N1

To analyze the amplitude of the N1, we ran a 2 (Modality; Audiovisual versus Auditory-only) x 2 (Group; Speech- versus Non-speech Mode) repeated measures ANOVA. N1 amplitude for audiovisual presentations (AVcongruent – V) was 1.08 μV smaller than for auditory-only presentations ($F(1,26) = 6.61$, $p < .017$). N1 amplitude did not differ across groups and N1 attenuation was not modulated by group (F-values < 1). The same ANOVA on the N1 latencies showed a 9.84 msec visually induced speeding up of the N1 ($F(1,26) = 22.62$, $p < .001$) that was alike in both groups ($F < 1$). There was a main effect of group ($F(1,26) = 5.86$, $p < .023$) as overall, the N1 peaked earlier in the non-speech- than speech group (at 114 vs 127 msec respectively).

7.4.2 - P2

For the P2, the 2 (Modality; Audiovisual versus Auditory-only) x 2 (Group; Speech- versus Non-speech Mode) repeated measures ANOVA on the amplitude showed a main effect of Modality ($F(1,26) = 7.70$, $p < .011$) and no main effect of group ($F < 1$). The interaction between Group and Modality approached significance ($F(1,26) = 3.39$, $p < .078$). As can be seen in Figure 7a, this finding reflects that visually induced P2 attenuation occurred in the speech- but not in the non-speech group. This was indeed confirmed by two separate t-tests between the A-only and AVcongruent – V P2 amplitudes ($t(13) = 3.32$, $p < .006$ for the speech group versus $t(13) = .65$, $p = .526$ for the non-speech group). The P2 latencies were alike for the audiovisual and auditory-only conditions ($F(1,26) = 2.67$, $p = .11$), and did not differ across groups ($F < 1$). There was no interaction between Group and Modality ($F(1,26) = 1.29$, $p = .267$).

The spatio-temporal dynamics of the AV interaction were further explored by conducting point-by-point two-tailed t-tests on the congruent audiovisual difference wave (AVcongruent – V – A) at each electrode in a 1 – 300 msec window. Relying on a procedure to minimize type I errors (Guthrie & Buchwald, 1991), AV interactions were considered significant when at least 12 consecutive points (i.e., 24 msec when the signal was re-sampled at 500 Hz) were significantly different from zero. This analysis allowed for detection of the earliest time where AV interactions occurred (see Figure 7aC). This analysis revealed reliable AV interactions for both groups within the 100 – 200 msec window, corresponding to the modulations of N1. For both groups, the effect was most prominent for the fronto-central electrodes. Additionally, only in the speech group, we observed later interactions corresponding to the P2 modulations at central-anterior locations.

7.4.3 - Audiovisual congruency

We also examined whether the ERPs showed an effect of stimulus congruency. As noted, when measured from sound onset, AV incongruency in our stimuli came too late (viz, > 270 msec) to be reflected in the P2 (and (in)congruency effects were actually not found at P2), so any effect of stimulus congruency should occur later than 270 msec.

We analyzed the data for Ab (AbVb versus AbVg) separately from Ag (AgVg versus AgVb) because in the speech mode, the AbVg stimuli possibly produced a fused /tadi/-McGurk-percept (McGurk & MacDonald, 1976) or a large lipread bias towards /tagi/ as demonstrated before (Vroomen & Stekelenburg, 2011). Both perceptual solutions to the AV incongruency indicate that the amount of perceived AV conflict is drastically reduced. In contrast, the AgVb stimuli are presumably perceived as a genuine conflict (/tabgi/ or /tagbi/). To analyze congruency effects we constructed difference waves by subtracting the AV congruent ERPs from the incongruent ones (see Figure 7b).

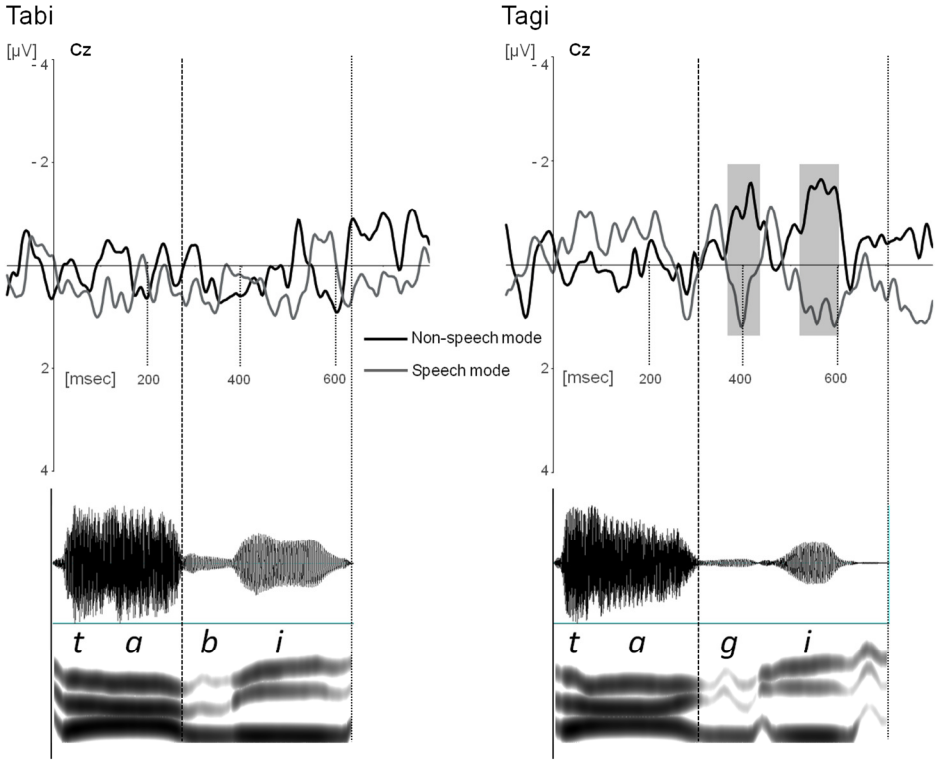

Figure 7b. *The speech- and non-speech modes' difference waves (AVI-V – AVC-V) of the audiovisual ERPs recorded at electrode Cz for /tabi/ and /tagi/. The SWS oscillograms and spectrograms are displayed to indicate the onset of AV incongruency at the second consonant and offset of the auditory signal. The grey areas indicate time-windows in which group differences were observed.*

As indicated in Figure 7b, the difference waves for the /tagi/ stimuli (AgVb - AgVg) showed two areas where the speech group potentially differed from the non-speech group; the area in between ~380 – 420 msec and the area in between ~520 – 600 msec. We therefore calculated the mean values of the difference waves for both areas and conducted two t-tests on the group differences. As anticipated, the /tagi/ difference waves for the speech-mode differed from the non-speech mode in both areas; ~380 – 420 msec (0.76 vs. -1.17; $t(26) = 2.15$, $p < .042$) and ~520 – 600 msec (0.88 vs. -1.35; $t(26) = 2.11$, $p < .045$).

7.5 - Discussion

The data encompasses four main findings; (1) the auditory N1 is attenuated and speeded-up by addition of the visual stimulus irrespective of whether the SWS stimuli were perceived as speech or not, (2) the auditory N1 peaked later in the speech- than non-speech group, (3) the P2 was only attenuated when the stimuli were perceived as speech and (4) congruency effects were observed after the P2 modulations and were restricted to the speech-mode.

For the N1, the results are in line with the notion that the N1 is modulated by a rather low-level prediction mechanism driven by the anticipatory visual motion that alerts the listener about when a sound is going to occur (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010) and is likely constituted via a direct route from visual to auditory brain areas (Arnal et al., 2009). In correspondence, it has been demonstrated that lipread speech increases detection of auditory speech when the sound is masked in noise (Grant & Seitz, 2000) but not if the speech is played backwards (Kim & Davis, 2004), presumably because the natural predictive value of the lipread signal is wiped out. Overall, these examples provide a rather compelling case to suggest that the natural temporal characteristics of the AV speech signal are not only integrated on a *different* level than phonetic and semantic features, but also *before* the stimuli are processed phonetically. The data additionally showed a more sluggish N1 (i.e., the N1 peaked later) for the speech- than non-speech group, presumably reflecting that speech processing is more strenuous than processing non-speech material.

Since phonetic AV binding is not reflected by the N1, the most prominent ERP component that presumably is sensitive to phonetic information is the P2. Although Stekelenburg and Vroomen (2007) already hypothesized that lipread induced P2 attenuations might reflect an effect of perceived phonetic stimulus congruency, it should be noted that this inference was considerably indirect as phonetic (in)congruency could not be disentangled from temporal (in)congruency because the authors used speech stimuli in which stimulus (in)congruency was already apparent at sound onset (i.e., /bi/

versus /fu/). Here, we obtained rather direct evidence that the P2 is indeed sensitive to phonetic information because only listeners that were aware of the speech-origin of the stimuli showed a lipread induced attenuation of P2. Given that both modality inputs were always phonetically and temporally congruent in the critical time window of 150 – 250 msec, it appears that the lipread induced P2 attenuations reflect phonetic binding between the auditory and visual speech signal that occurs independent of perceiving AV congruence. As noted however, it is unlikely that the P2 is only sensitive to phonetic information because there are many active neuronal processes in this time-window so AV incongruency, if apparent at sound onset, could potentially have an additional super- or supra additive effect on the P2 peak.

The second dissociation between the speech-and non-speech mode was found in the effect that congruency had on the brain potentials. These group differences, which are likely caused by the perceived AV conflict in the AgVb stimulus in the speech-mode only, further underscore the notion that AV SWS is processed differently when listeners are aware of the speech-origin of the stimuli (Tuomainen et al., 2005; Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011). As noted, processing of SWS sounds as speech differentiates itself from processing the same sounds as non-speech in STS (Benson et al., 2006; Lee & Noppeney, 2011; Möttönen et al., 2006). Interestingly, Arnal et al. (2009) proposed that STS is connected to motion areas and auditory cortex and several loops between these structures (taking over 500 msec and starting ~20 msec after the N1 is generated) are needed in order to tune the system towards stable percept of AV (in)congruence. Admittedly, the observed group differences at ~380 – 420 and ~520 – 600 msec did not take more than 500 msec to get constituted because AV incongruence did not become apparent before 270 msec. Nevertheless, there is a possibility that, the group difference at ~380 – 420 msec could reflect a perceptual outcome of a relatively early interaction loop between the involved brain areas that is further processed and tuned in another loop, yielding a stable and final incongruent percept at ~520 – 600 msec. STS presumably has a mediating key role in this process since it is argued that when visual predictions are violated by incongruent auditory input, there is a coupling between low beta- (14-15 Hz) and high gamma activity within STS, likely reflecting that stimulus incongruency generates prediction errors in auditory and visual cortices which get up-dated in STS (Arnal, Wyart, & Giraud, 2011).

Taken together, our data suggest that the underlying time-course of AV speech perception reflects at least three subsequent levels of integration in which the natural temporal characteristics of an audiovisual speech signal are integrated (at N1) before the modality inputs are bound phonetically (at P2) and stimulus congruency is processed and perceptually finalized.

Chapter 8

Discussion

8.1 - Summary of the results

The experimental results throughout the literature, including those described in this thesis, provide a solid platform to suggest that lipread induced phonetic recalibration is not a coincidental finding constituted by a particular method, experimental design or set of stimuli. Instead, the unimodal aftereffects are rather universal in the sense that they are reported in a large body of literature with the same essential message; a speech context containing information about ambiguous auditory speech input can re-adjust the perceptual system on a longer term basis.

This thesis has sought to show new insights in the recalibration aftereffects and its underlying mechanisms. The first data chapter was set-up to re-investigate the notion that recalibration effects are rather short-lived (**Chapter 2**). Although short-lived, phonetic recalibration is apparently very robust as it is not influenced by extreme violations of the natural characteristics of the speech signal as long as the auditory and lipread inputs are bound on a phonetic level (**Chapter 3**), which possibly occurs at around 200 msec after sound-onset (**Chapter 7**). Recalibration is bi-directional in nature, in close correspondence with other audiovisual illusions such as the ventriloquist effect (**Chapter 4**). Phonetic recalibration seems, as a first approximation, independent of working memory (**Chapter 5**) and impairments in reading (or difficulties with learning to read) and auditory-only speech identification (**Chapter 6**).

8.2 - Towards a conclusion

The literature overview provided in the first chapter of this thesis underscores the consensus that the speech system is flexible, dynamic and capable of making perceptual adaptations (see Samuel, 2011; Vroomen & Baart, 2011, for reviews). Given that phonetic recalibration of the system is bi-directional and rather robust against experimental manipulations, the following conclusion can be drawn:

Recalibration effects can be taken as an indication that the speech system is dynamic because longer-term perceptual auditory adjustments can be flexibly constituted by lipread or lexical speech context.

In what follows, this conclusion is further specified based on available literature and future directions and the mechanism that possibly underlies recalibration are discussed (8.3).

8.2.1 - Flexibility

Repeated exposure to an ambiguous speech sound in between two categories combined with unambiguous lipread or lexical speech context results in a transient shift in the auditory phoneme boundary. That is, a previously ambiguously perceived sound is perceptually categorized as a member of a particular speech category based on the previously repeatedly delivered context. Apparently, the speech system is flexibly adjusted to solve a conflict between auditory input and relevant context information. However, the story is more complicated as exemplified by a study by Kraljic, Samuel and Brennan (2008) in which listeners were presented with an ambiguous speech token in between /s/ and /sh/ embedded in words. As expected, these stimuli induced lexically-driven recalibration effects, but interestingly, this process was blocked whenever listeners had previously heard good standard pronunciations of the sound from the same speaker or when the speaker placed a pen in the mouth whenever the ambiguous sound was heard. It thus appears that combining ambiguous speech with a disambiguating lexical context is not enough to induce recalibration; the listener should additionally be under the impression that the ambiguous sound is not created by accidental mispronunciation or an obvious obstruction in the airflow. Moreover, when ambiguous /s/-/sh/ sounds can be attributed to a speaker's dialect, known to the listener, lexically-driven recalibration is not observed. In contrast, when the same sound is attributed to individual speech characteristics of the speaker, recalibration will occur (Kraljic, Brennan, et al., 2008), in line with a study of Eisner and McQueen (2005, Experiment 3).

It thus appears that recalibration is countered by additional information available to the listener that can explain why the sound is ambiguous (a pen in the mouth of the speaker or a particular dialect) and is speaker specific.

Why would a low-level process like recalibration be constrained by these 'high level' cognitive and socio-linguistic factors? In order to answer this question, one needs to consider that the primary function of human speech perception is to perceive a specific message. It is plausible to assume that in daily life situations, there is not always a need to adapt the system in order to accurately perceive the message as intended. Just as an example, an ambiguously perceived b/d sound embedded in the sentence "Watch out for that treacherous bog/dog" is most likely interpreted as a /d/ rather than a /b/ when the listener finds himself in a rescue center for neglected animals and sound ambiguity was created by a dog barking in the background.

In this case, the environmental cues leave little room for a /b/ and it seems rather far-fetched to assume that repeated exposure to this specific situation (as is the case in the experimental situations created in the laboratory) would actually elicit an adaptation in the speech system.

As for speaker specificity, it can be argued that it is unnecessary to generalize an adaptation of the system towards other speakers than the one you are currently interacting with because inter-speaker variability and varying noise levels might well require speaker- and situation-specific adaptations. Similarly, it is presumably unnecessary to adjust the system towards a particular dialect the listener is already familiar with, because the context can explain sound ambiguity. Incorporating these inferences in the conclusion would yield:

Recalibration effects can be taken as an indication that the speech system is dynamic because (1) longer-term perceptual auditory adjustments can be flexibly constituted by lipread or lexical speech context and (2) corrective adjustments are only made when they are needed for accurate future perception.

However, this notion introduces a rather intangible suggestion that the brain has to process all the relevant features of the specific context, make a decision about its quality and feed that back into the speech system that has to come-up with a timely recalibration procedure.

A possibility though, is that the contextual cues most relevant to auditory speech perception, namely the preceding lip movements and the lexicon, are processed early (enough) to prepare for recalibration. Feedback generated by processing additional features of the context could then prevent recalibration from developing as its corrective

function has become redundant. This however implies that there should at least be evidence that lipread speech and lexical context are indeed processed early.

8.2.2 - Recalibration as a default state?

One way to improve accuracy of speech perception is to rely on multiple modalities that provide the brain with shared information about specific stimulus properties such as spatial location and duration (Lewkowicz & Kraebel, 2004). For speech perception, the second modality involved in the process is vision. The lipread speech stream shares information about temporal properties, source, location and speech sound identity with the auditory input and it seems likely that the relative reliability of the lipread information increases whenever the auditory reliability decreases, in line with the principle of 'inverse effectiveness' (e.g., Stein & Meredith, 1993). Because it is conceivable that the speech system derives its accuracy partially from corrective lipread input on an on-line basis (e.g., Sumby & Pollack, 1954) it may come as no surprise that there is evidence that suggests a coupling of the two inputs in the brain. The most prominent brain area that is reported to be highly involved in audiovisual speech perception is the Superior Temporal Sulcus, or STS, located in the superior temporal lobe (e.g., Arnal et al., 2009; Arnal et al., 2011; Calvert et al., 1999; Calvert & Campbell, 2003; Calvert, Campbell, & Brammer, 2000; Möttönen, Schürmann, & Sams, 2004).

Rather than that the accuracy of the speech system is derived from the auditory signal alone, it thus seems to co-depend on a functional link between acoustic and lipread inputs that possibly is initiated at around ~200 msec after sound onset (see Chapter 7).

It may also seem likely that lexical context is functionally linked to relative early auditory speech processes, but the evidence is mainly derived from behavioral studies that showed that lexical speech context is capable of inducing so-called 'selective speech adaptation' effects (Samuel, 1997, 2001). Since selective speech adaptation is argued to depend on low-level acoustic factors (e.g., Sawusch, 1977), it is conceivable that the effect arises early during auditory processing. The possibility that lexical information penetrates early auditory processes is corroborated by interactive approaches that assume lexical influences on pre-lexical processing (McClelland, Mirman, & Holt, 2006) and the finding that a lexically induced change in perceived sound identity reduces the Mismatch Negativity (MMN) at ~200 msec after stimulus onset (van Linden et al., 2007).

Assuming that both lipread and lexical speech contexts have an early access to auditory processes, it might well be that the fundaments of phonetic recalibration

aftereffects are constituted relatively early during processing. In line with this notion, it has been argued that the on-line perceptual biases as found for both lipread information (e.g., McGurk & MacDonald, 1976; Sumby & Pollack, 1954) and lexical context (e.g., Ganong, 1980) are part of the same mechanism (that is responsible for maintaining perceptual coherence in speech) as recalibration effects (van Linden, 2007, Chapter 9).

Interestingly, recalibration is not restricted to the speech domain as similar aftereffects are observed for perceived location (e.g., Bertelson, 1999; Radeau & Bertelson, 1974) and perceived audiovisual synchronicity (e.g., Keetels, Stekelenburg, & Vroomen, 2007; Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Vroomen & Keetels, 2006). It thus seems likely that corrective recalibration effects are the rule rather than the exception, providing support for the idea that speech recalibration is initiated automatically (e.g., Baart & Vroomen, 2010b; McQueen, Norris, et al., 2006) and might reflect a default mechanism in the brain that is triggered whenever there is a perceptual speech-, location- or temporal conflict. The conclusion could then be formulated as follows:

Recalibration effects can be taken as an indication that the speech system is dynamic because (1) longer-term perceptual auditory adjustments can be flexibly constituted by lipread or lexical speech context and (2) default corrective adjustments to the system can be blocked when they are not needed for accurate future perception.

When formulating any conclusion about speech related processes, it is inevitable to consider a well discussed and controversial notion about speech, namely, that speech perception is special.

8.2.3 - Is speech special?

Human speech has no natural equivalent in terms of acoustical complexity, semantics, syntax and phonetics. As such, it is likely that speech perception (partially) relies on unique processes. Speech processing thus differs from processing non-speech and therefore, speech is special.

If only it was that simple. The problem with this inference is that it is rather circular. If one wants to argue that an elephant is a special animal, it seems rather elusive to state that the elephant has a set of unique properties such as a trunk, a pair of tusks and big ears and thus, an elephant is special. This might be considered true, but then all animals (and all multisensory input) are special because of their unique features and properties.

As mentioned, it is however conceivable that speech processing, to a certain extent, relies on specific processes. First and foremost, correct perception of the acoustic speech waves requires a discriminative and precise auditory system. However, being precise does not imply that all acoustic variations introduced by the characteristics of the speaker or the transient variations in levels of environmental background noise should be processed in detail. Rather, being precise implies that the correct auditory speech alternative is accurately perceived. So a system should therefore be better in discriminating between sounds that belong to a different category, and thus have a different meaning, than between sounds that belong to the same category because acoustic differences within a category are not important for the communicative value of the message. Indeed, humans are better in discriminating between speech sounds that belong to different phoneme categories, such as /b/ or /p/, than they are at discriminating between physically equally-different sounds that fall in the same category (i.e., so-called categorical perception or CP, e.g., Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Harnad, 1987; Repp, Healy, & Crowder, 1979). It should be noted though, that CP is also observed with non-speech sounds (Cutting & Rosner, 1974; Jusczyk, Rosner, Cutting, Foard, & Smith, 1977) so better auditory between- than within category discrimination is not restricted to speech. In addition Samuel (1977) demonstrated that after extensive training (i.e., multiple sessions a week for several weeks) listeners are quite able to discriminate between within-category speech sounds.

Importantly, one should be aware of the fact that effective communication through speech does not solely depend on the listener's brain that has to deal with the input. Accurate speech production is equally important to get the message across as intended and successful communication obviously falters if either one of the tasks involved (i.e., producing and perceiving speech) is compromised.

Accuracy and efficiency of such a complex combination of motor and perceptual tasks that might be indicative of the special nature of speech are, according to the 'Motor theory of speech' (Galantucci, Fowler, & Turvey, 2006; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985), constituted through a systematic biological link between brain areas involved in (audiovisual) perception and production of speech (Liberman & Mattingly, 1985). The existence of such a link is corroborated by studies on so-called mirror neurons, that presumably aid understanding and action and mediate imitation (Rizzolatti & Craighero, 2004), that have suggested that the human homologue of the macaque inferior premotor cortex, Broca's area, is involved in the production of speech and is activated during silent lipreading (Campbell et al., 2001) and when perceiving auditory speech (Wilson et al., 2004).

Another piece of evidence to support the existence of speech specific perceptual processes is provided by studies using sine-wave speech (SWS) as the highly artificial sounds can be perceived as speech but only if the listener is informed about the speech-origin of the material (Remez et al., 1981). Since naïve listeners are able to discriminate between SWS alternatives based on acoustic properties alone, it seems conceivable that there are speech specific phonetic processes that run in parallel with processing acoustic factors as already proposed before the first report on SWS (Fujisaki & Kawashima, 1969, 1970; Pisoni, 1973). In close correspondence, it has been demonstrated that phonetic audiovisual integration of SWS sounds only occurs when SWS is perceived as speech (Eskelund et al., 2011; Tuomainen et al., 2005; Vroomen & Baart, 2009a; Vroomen & Stekelenburg, 2011) although temporal binding of both modality inputs is independent of the speech/non-speech interpretation of the signal (Vroomen & Stekelenburg, 2011) and the so-called detection advantage (i.e., a visually induced auditory detection benefit) is also unaffected by prior knowledge about the SWS sounds (Eskelund et al., 2011).

Interestingly, functional Magnetic Resonance Imaging (fMRI) studies incorporating SWS have again underscored the crucial role of the Superior Temporal Sulcus (STS) in speech perception because activity in STS increases when SWS is perceived as speech (Benson et al., 2006; Möttönen et al., 2006) and possibly, during audiovisual integration of SWS stimuli, sensitivity to stimulus intelligibility and linguistic access are reflected in anterior regions of STS (Lee & Noppeney, 2011).

Overall, it is quite clear that there are speech specific properties that elicit certain perceptual processes that can be measured in the brain as well as on a behavioral level. Although it could be argued that it is just a case of semantics, I would like to state that speech is not special but the speech system as such has a specialized task.

When does a conclusion become final? My educated guess would be ‘never’ since there is always room for adjustments and specifications as research progresses (see 8.3 for suggestions). However, at this point I feel it is appropriate to conclude:

Recalibration effects can be taken as an indication that the specialized speech system is dynamic because (1) longer-term perceptual auditory adjustments can be flexibly constituted by lipread or lexical speech context and (2) default corrective adjustments to the system can be blocked when they are not needed for accurate future perception.

8.3 - Future directions and underlying mechanisms

8.3.1 - Lipread versus lexically induced recalibration; future directions

Since phonetic recalibration is measured in an auditory only task, it is indicative of auditory perceptual changes. There seem to be two prominent possibilities that could account for the observed effects;

(1) Recalibration induces a shift of the perceived phoneme boundary towards the context.

(2) Recalibration induces a widening of the phoneme category to the extent that the ambiguous sound gets included in the specific category.

For lexical recalibration, it appears that the second option is most likely as it has been demonstrated that recalibration is not restricted to the ambiguous sound only, instead, the phoneme category seems widened because after exposure, the auditory continuum is almost entirely rated more in accordance with the lexically defined category than before exposure (e.g., Kraljic & Samuel, 2005; Norris et al., 2003). In contrast, the experiments described in this thesis followed the standard paradigm used in lipread recalibration and used only ambiguous sounds during the test. Although lipread recalibration is most prominently obtained with the most ambiguous sound of the three test-stimuli, it is currently unclear whether lipread recalibration induces similar perceptual widening of the phoneme category as seems to be the case for lexical recalibration.

Given that there are crucial differences between lipread and lexical recalibration, it could well be that the auditory aftereffects reflect different auditory adjustments. Firstly, it is appropriate to note that lipread input is bottom-up visual information that is usually perceived in synchrony with the auditory stream, although it actually often precedes the acoustic signal (Chandrasekaran et al., 2009). Lexical information, in contrast, exerts a top-down influence on auditory processing, and is argued to become important as the word is being recognized (e.g., Samuel & Pitt, 2003).

Secondly, lexical information is capable of inducing selective speech adaptation while lipread information is not. When for instance a white noise burst was placed in the lexical context of ‘alpha?et’, listeners heard the sound as a /b/ and the sound additionally induced selective adaptation as could be expected from a canonical /b/ (i.e., more /d/-responses on the continuum tokens, see Samuel, 1997). Likewise, an ambiguous mixture of /s/ and /sh/ placed in the context of ‘aboli?’ was perceived as /sh/ and again, selective adaptation was observed (Samuel, 2001). For lipread information however, a successful McGurk-illusion (listeners perceive a /d/ when lipread /g/ is

paired with an auditory /b/) does not yield selective adaptation in concordance with a /d/ percept. Instead, selective adaptation effects of the McGurk-/d/ (obtained with auditory /b/) were the same as selective adaptation effects obtained with a genuine /b/ (e.g., Roberts & Summerfield, 1981). Thirdly, compensation for coarticulation (Repp & Mann, 1981) can be obtained through lexical speech context, as a lexically disambiguated /s/–/ʃ/ sound produces context effects on identifying members of a /t/-/k/ continuum (Elman & McClelland, 1988), but cannot be obtained by lipread context (Holt, Stephens, & Lotto, 2005; Vroomen & de Gelder, 2001).

As noted earlier, there are also differences in lipread and lexical recalibration; lexically induced recalibration is reported to be long lasting (Eisner & McQueen, 2006; Kraljic & Samuel, 2005) whereas lipread induced recalibration is short lived (e.g., van Linden & Vroomen, 2007; Vroomen & Baart, 2009b). Additionally, for lexically induced recalibration, it has also been demonstrated that recalibration is depending on idiosyncratic features of the speaker (Eisner & McQueen, 2005; Kraljic, Brennan, et al., 2008) but can generalize to words outside the original training set (McQueen, Cutler, et al., 2006) and across syllabic positions (Jesse & McQueen, 2011). For lipread induced recalibration, these possibilities have not been investigated.

A rather straightforward suggestion for future research is to determine whether the lipread induced recalibration aftereffects follow the same pattern of generalizations as the lexically induced aftereffects.

8.3.2 - Underlying mechanism

In the lexical case, phonetic recalibration can be explained by interactive models such as TRACE (McClelland & Elman, 1986). In TRACE (and TRACE II), it is proposed that a large number of processing units are organized into three levels, namely, features, phonemes and words. Units on the *same* and *different* levels interact with each other through bidirectional excitatory and inhibitory connections. To be more precise, it is proposed that units on the *same* level influence each other exclusively through inhibitory connections whereas *between* level connections are always excitatory. This results in a dynamic process in which each unit continuously updates its own activation based on activation of other units to which it is connected. It would then be conceivable that phonetic recalibration is constituted through strengthening and re-tuning of the connective pathways between the three levels. If the ambiguous auditory input is perceived again without lexical information (i.e., during the test) it would then be processed via the re-tuned feature-to-phoneme connections, yielding the same perceptual solution as before (Mirman et al., 2006).

Although the lipread input is not incorporated in this model, it seems conceivable that there could be an additional layer concerned with processing visual features (needed for lipreading) that is connected with the auditory phoneme layer as described in TRACE. This visual-feature-to-phoneme connection could explain on-line effects of lipreading on auditory perception (such as the McGurk-illusion) as well as lipread induced recalibration because, in analogy with lexical context, visual feature-to-phoneme connections can be re-tuned and strengthened by the lipread context, following a Hebbian learning principle (Hebb, 1949; Mirman et al., 2006).

References

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*(9), 839-843.

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research, 183*(3), 399-404.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Washington DC: American Psychiatric Association.

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., & Sams, M. (2009). The role of visual attention in audiovisual speech perception. *Speech Communication, 51*, 184-193.

Anstis, S. (1986). Motion perception in the frontal plane: Sensory aspects. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, chap. 16). New York: Wiley.

Anstis, S., Verstraten, F. A. J., & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences, 2*, 111-117.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience, 29*(43), 13445-13453.

Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience, 14*(6), 797-U164.

Baart, M., & Vroomen, J. (2010a). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters, 471*(2), 100-103.

Baart, M., & Vroomen, J. (2010b). Phonetic recalibration does not depend on working memory. *Experimental Brain Research, 203*(3), 575-582.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47-89). New-York: Academic Press.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). New York: Cambridge University Press.

Benson, R. R., Richardson, M., Whalen, D. H., & Lai, S. (2006). Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. *Neuroimage, 31*(1), 342-353.

Bermant, R. I., & Welch, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual & Motor Skills, 42*(43), 487-493.

Bertelson, P. (1999). Ventriloquism: A case of cross-modal grouping. In G. Aschersleben, T. Bachmann & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347-362). Amsterdam: Elsevier.

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*(3), 482-489.

Bertelson, P., Frissen, I., Vroomen, J., & De Gelder, B. (2006). The aftereffects of ventriloquism: Patterns of spatial generalization. *Perception & Psychophysics, 68*(3), 428-436.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics, 29*(6), 578-584.

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

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62*, 321 - 332.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20*(8), 2225-2234.

Blakemore, C., & Sutton, P. (1969). Size adaptation: a new aftereffect. *Science, 166*(902), 245-247.

Blomert, L., & Mitterer, H. (2004). The fragile nature of the speech-perception deficit in dyslexia: Natural vs. synthetic speech. *Brain and Language, 89*(1), 21 - 26.

Boersma, P., & Weenink, K. (2005). Praat: doing phonetics by computer, Retrieved from http://www.fon.hum.uva.nl/praat.

Bogliotti, C., Serniclaes, W., Messaoud-Galusi, S., & Sprenger-Charolles, L. (2008). Discrimination of speech sounds by children with dyslexia: Comparisons with chronological age and reading level controls. *Journal of Experimental Child Psychology*, *101*, 137-155.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 30*(3), 445-463.

Brus, B. T., & Voeten, M. J. M. (1997). *Een-minuut-test*. Amsterdam: Pearson.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*(17), 2213-2218.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16*(5), 805-816.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*(12), 2619-2623.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593-596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15*(1), 57-70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649-657.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1493), 1001-1010.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Brain Research, Cognitive Brain Research, 12*(2), 233-243.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology, 5*(7), e1000436.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology, 113*(4), 495-506.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries *Proceedings of Interspeech 2008*: Brisbane, Australia.

Cutting, J. E., & Rosner, B. S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics, 16*(3), 564 - 570.

de Gelder, B., & Vroomen, J. (1998). Impaired speech perception in poor readers: Evidence from hearing and speech reading. *Brain and Language, 64*, 269–281.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*, 187-203.

Diehl, R. L. (1981). Feature detectors for speech: a critical reappraisal. *Psychological Bulletin, 89*(1), 1-18.

Diehl, R. L., Elman, J. L., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Human Perception & Performance, 4*(4), 599-609.

Diehl, R. L., Lang, M., & Parker, E. M. (1980). A further parallel between selective adaptation and contrast. *Journal of Experimental Psychology: Human Perception & Performance, 6*(1), 24-44.

Ebbinghaus, M. (1885). *Über das Gedächtnis*. K. Buehler, Leipzig.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*, 99-109.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*, 303 - 306.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224-238.

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

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*, 143-165.

Epstein, W. (1975). Recalibration by pairing: A process of perceptual learning. *Perception, 4*, 59-72.

Erber, N. P. (1974). Auditory-visual perception of speech: A survey. In H. B. Nielsen & E. Kampp (Eds.), *Visual and audio-visual perception of speech*. Stockholm, Sweden: Almquist & Wiksell.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162-169.

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: dissociating identification and detection. *Experimental Brain Research, 208*(3), 447-457.

Fairhall, S. L., & Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *European Journal of Neuroscience, 29*, 1247-1257.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science, 322*(5903), 970-973.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Early auditory–visual interactions in human cortex during nonredundant target identification. . *Brain Research, Cognitive Brain Research, 14*, 20–30.

Fujisaki, H., & Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual report of the Engineering Research Institute* (Vol. 28, pp. 67 - 73). Tokyo: University of Tokyo, Faculty of Engineering.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual report of the Engineering Research Institute* (Vol. 29, pp. 207 - 214). Tokyo: University of Tokyo, Faculty of Engineering.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), *361-377*.

Ganong, W. F. (1978). The selective adaptation effects of burst-cued stops. *Perception & Psychophysics, 24*(1), 71-83.

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

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*(5), 473-490.

Gibson, J. J. (1933). Adaptation, after-effects and contrast in the perception of curved lines. *Journal of Experimental Psychology, 18*, 1–31.

Godfrey, J. J., Syrdal-Lasky, K., Millay, K. K., & Knox, C. M. (1981). Performance of dyslexic children on speech perception tests. *Journal of Experimental Child Psychology, 32*(3), 401-424.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*(3), 1197-1208.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*(4), 468-484.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology, 28*(2), 240-244.

Guttorm, T. K., Leppanen, P. H. T., Poikkeus, A. M., Eklund, K. M., Lyytinen, P., & Lyytinen, H. (2005). Brain event-related potentials (ERPs) measured at birth predict later language development in children with and without familial risk for dyslexia. *Cortex, 41*(3), 291-303.

Guttorm, T. K., Leppanen, P. H. T., Tolvanen, A., & Lyytinen, H. (2003). Event-related potentials in newborns with and without familial risk for dyslexia: principal component analysis reveals differences between the groups. *Journal of Neural Transmission, 110*(9), 1059-1074.

Harnad, S. (Ed.). (1987). *Categorical perception: the groundwork of cognition*. Cambridge: Cambridge University Press.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Held, R. (1965). Plasticity in sensory-motor systems. *Scientific America, 213*(5), 84-94.

Hidaka, S., Manaka, Y., Teramoto, W., Sugita, Y., Miyauchi, R., Gyoba, J., et al. (2009). Alternation of Sound Location Induces Visual Motion Perception of a Static Object. *PLoS One, 4*(12, e8188), 1-6.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77 - 99). Potomac, MD: Erlbaum.

Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision - A comparison using Event-related potentials. *Language and Cognitive Processes, 5*(4), 281-312.

Holt, L. L., Stephens, J. D. W., & Lotto, A. J. (2005). A critical evaluation of visually moderated phonetic context effects. *Perception & Psychophysics, 67*(6), 1102-1112.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review, 18*, 943-950.

Jonides, J., Schumacher, E. H., Smith, E. E., Koeppe, R. A., Awh, E., Reuter-Lorenz, P. A., Marhuetz, C., & Willis, C. R. (1998). The Role of Parietal Cortex in Verbal Working Memory. *The Journal of Neuroscience, 18*(13), 5026 - 5034.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F., & Smith, L. B. (1977). Categorical perception of non-speech sounds by 2-month-old infants. *Perception & Psychophysics, 21*(1), 50 - 54.

Keetels, M., Stekelenburg, J., & Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: evidence from temporal ventriloquism. *Experimental Brain Research, 180*(3), 449-456.

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *The Journal of Neuroscience, 31*, 1715 - 1720.

Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2008). One sound, two percepts: Predicting future speech perception from brain activation during audiovisual exposure. *Neuroimage, 41, Supplement 1*, S112.

Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2011). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *Neuroimage, 57*(4), 1601-1607.

Kim, J., & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication, 44*(1-4), 19-30.

Klemm, O. (1909). Localisation von Sinneneindrücken bei disparaten Nebenreizen. *Psychologische Studien, 5*, 73-161.

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research, Cognitive Brain Research, 18*(1), 65-75.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition, 107*(1), 54-81.

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

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262-268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*, 1-15.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: how listeners adjust to speaker variability. *Psychological Science, 19*(4), 332-338.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138-1141.

Lang, H., Nyrke, T., Ek, M., Aaltonen, O., Raimo, I., & Näätänen, R. (1990). Pitch discrimination performance and auditory event-related potentials. In C. M. H.

Brunia, A. W. K. Gaillard, A. Kok, G. Mulder & M. N. Verbaten (Eds.), *Psychophysiological Brain Research, vol. 1* (pp. 294–298). Tilburg: Tilburg University Press.

Lee, H., & Noppeney, U. (2011). Physical and Perceptual Factors Shape the Neural Mechanisms That Integrate Audiovisual Signals in Speech Comprehension. *Journal of Neuroscience, 31*(31), 11338-11350.

Leppanen, P. H. T., Pihko, E., Eklund, K. M., & Lyytinen, H. (1999). Cortical responses of infants with and without a genetic risk for dyslexia: II. Group effects. *Neuroreport, 10*(5), 969-973.

Lewkowicz, D. J., & Kraebel, K. S. (2004). The value of multisensory redundancy in the development of intersensory perception. In G. Calvert, C. Spence & B. Stein (Eds.), *The handbook of multisensory processes* (pp. 655 - 678). Cambridge, MA: The MIT press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431-461.

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

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777-1788.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge: The MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(8), 363-369.

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

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science, 30*, 1113-1126.

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical-prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language, 61*, 1-18.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The Dynamic Nature of Speech Perception. *Language and Speech, 49*(1), 101-112.

McQueen, J. M., Tyler, M., & Cutler, A. (in press). Lexical retuning of children's

speech perception: Evidence for knowledge about words' component sounds. *Language Learning and Development*.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*(6), 958-965.

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. [Proceedings Paper]. *Clinical Linguistics & Phonetics, 20*(7-8), 621-630.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Brain Research, Cognitive Brain Research, 17*(1), 154-163.

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage, 30*(2), 563-569.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Research, Cognitive Brain Research, 13*(3), 417-425.

Möttönen, R., Schürmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neuroscience Letters, 363*(2), 112-115.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238.

Näätänen, R. (1992). *Attention and Brain Function*: Hillsdale, NJ: Erlbaum.

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent. *Psychophysiology, 38*, 1-21.

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect in evoked potential reinterpreted. *Acta Psychologica, 42*, 313-329.

Näätänen, R., Paavilainen, P., Tiitinen, H., Jiang, D., & Alho, K. (1993). Attention and mismatch negativity. *Psychophysiology, 30*, 436-450.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25*(2), 333-338.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development, 22*, 237-247.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191-196.

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, I. P., Kujala, T., et al. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3 T. *Neuroimage, 29*(3), 797-807.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Mottonen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport, 16*(2), 125-128.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*, 253 - 260.

Radeau, M., & Bertelson, P. (1974). The after-effects of ventriloquism. *The Quarterly Journal of Experimental Psychology, 26*(1), 63-71.

Radeau, M., & Bertelson, P. (1976). The effect of a textured visual field on modality dominance in a ventriloquism situation. *Perception & Psychophysics, 20*, 227-235.

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception & Psychophysics, 22*(2), 137-146.

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research, 49*(1), 17-22.

Ramirez, J., & Mann, V. (2005). Using auditory-visual speech to probe the basis of noise-impaired consonant-vowel perception in dyslexia and auditory neuropathy. *Journal of the Acoustical Society of America, 118*(2), 1122-1133.

Raven, J., Raven, J., & Court, J. H. (1998). *Raven manual: Standard progressive matrices.* Oxford, England: Oxford Psychologists Press.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947-949.

Repp, B. H., Healy, A. F., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception & Performance, 5*(1), 129-145.

Repp, B. H., & Mann, V. A. (1981). Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America, 69*, 1154-1163.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169-192.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics, 30*(4), 309-314.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*, 347-357.

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America, 95*(6), 3658-3661.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127*(1), 141-145.

Samuel, A. G. (1977). The effect of discrimination training on speech perception: Noncategorical perception *Perception & Psychophysics, 22*(4), 321-330.

Samuel, A. G. (1986). Red herring detectors and speech perception: in defense of selective adaptation. *Cognitive Psychology, 18*(4), 452-499.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*(2), 97-127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science, 12*(4), 348-351.

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology, 62*, 49 - 72.

Samuel, A. G., & Kat, D. (1996). Early Levels of Analysis of Speech. *Journal of Experimental Psychology: Human Perception & Performance, 22*(3), 676-694.

Samuel, A. G., & Kat, D. (1998). Adaptation is automatic. *Perception & Psychophysics, 60*(3), 503-510.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48*, 416–434.

Sawusch, J. R. (1977). Peripheral and central processes in selective adaptation of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 62*(3), 738-750.

Shigeno, S. (2002). Anchoring effects in audiovisual speech perception. *Journal of the Acoustical Society of America, 111*(6), 2853-2861.

Sjerps, M., & McQueen, J. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 36*(1), 195-211.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage, 25*(1), 76-89.

Smith, E. E., Jonides, J., & Koeppe, R. A. (1996). Dissociating verbal and spatial working memory using PET. *Cerebral Cortex, 6*(1), 11-20.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition, 92*(3), B13-23.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge: MIT-press.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964-1973.

Stevens, M. (2007). *Perceptual adaptation to phonological differences between language varieties.* University of Gent, Gent.

Stratton, G. M. (1896). Some preliminary experiments on vision without inversion of the retinal image. *Psychological Review*, 611-617.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerebral Cortex, 17*(3), 679-690.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*(3), 850-855.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457-472.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*(1), B13-22.

Van den Bos, K. P., Lutje Spelberg, H. C., Scheepsma, A. J. M., & De Vries, J. R. (1999). *De Klepel: Pseudowoordentest.* Amsterdam: Harcourt Test Publishers.

van Linden, S. (2007). *Recalibration of auditory phoneme perception by lipread and lexical information.* Tilburg University, Ridderprint BV, Ridderkerk.

van Linden, S., Stekelenburg, J. J., Tuomainen, J., & Vroomen, J. (2007). Lexical effects on auditory speech perception: An electrophysiological study. *Neuroscience Letters, 420*(1), 49-52.

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception & Performance, 33*(6), 1483-1494.

van Linden, S., & Vroomen, J. (2008). Audiovisual speech recalibration in children. *Journal of Child Language, 35*(4), 809-822.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181-1186.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598-607.

Vandermosten, M., Boets, B., Luts, H., Poelmans, H., Golestani, N., Wouters, J., et al. (2010). Adults with dyslexia are impaired in categorizing speech and non-speech sounds on the basis of temporal cues. *Proceedings of the National Academy of Sciences USA, 107*(23), 10389-10394.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): what have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*(1), 2-40.

von Helmholz, H. (1866). *Treatise on physiological optics*. vol. III, 3rd edition (translated by J. P. C. Southall 1925 *Optical Society America*. Section 26, reprinted New York: Dover Publications, 1962.

Vroomen, J., & Baart, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition, 110*(2), 254-259.

Vroomen, J., & Baart, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a twenty-four hours delay. *Language and Speech, 52*, 341-350.

Vroomen, J., & Baart, M. (2011). Phonetic recalibration in audiovisual speech. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 363-379). Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63*(4), 651-659.

Vroomen, J., & de Gelder, B. (2001). Lipreading and the compensation for coarticulation mechanism. *Language and Cognitive Processes, 16*, 661-672.

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective, & Behavioral Neuroscience, 1*(4), 382-387.

Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1063-1071.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual Anticipatory Information Modulates Multisensory Interactions of Artificial Audiovisual Stimuli. *Journal of Cognitive Neuroscience*.

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), 75-83.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45*(3), 572-577.

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*, 55-61.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech & Hearing Research, 20*, 130-145.

Wallach, H. (1968). Informational discrepancy as a basis of perceptual adaptation. In S. J. Freeman (Ed.), *The neuropsychology of spatially oriented behaviour* (pp. 209-230). Dorsey: Homewood, IL.

Welch, R. B., & Warren, D. H. (1986). Intersensory interactions. In K. R. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 1-36): Wiley.

Werker, J. F., & Tees, R. C. (1987). Speech perception in severely disabled and average reading children. *Canadian Journal of Experimental psychology, 41*, 48 - 61.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*(7), 701-702.

Winkler, I., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., et al. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology, 36*, 638-642.

Nederlandse samenvatting
Summary in Dutch

De waarneming van spraak is geen puur auditief proces maar is mede afhankelijk van het lexicon dat de luisteraar tot zijn beschikking heeft en het zien van de mondbewegingen van de spreker (hier aangeduid als 'liplezen'). Wanneer bijvoorbeeld een onduidelijke klank tussen een /b/ en een /d/ in wordt uitgesproken in de zin "Kunt u mij misschien de b/doter aangeven?", zal de luisteraar deze klank horen als een 'b' aangezien het woord "boter" een bestaand woord is en "doter" niet. Wanneer een luisteraar dezelfde klank (in het vervolg aangeduid als /A?/ dat staat voor 'auditief ambigue signaal') te horen krijgt terwijl de mond van de spreken een 'b' articuleert, zal de klank wederom als een 'b' worden waargenomen. Dit gebeurt doordat het auditieve en visuele signaal worden geïntegreerd tot één percept (audiovisuele integratie). Logischerwijs levert hetzelfde spraakgeluid in de zin "Ik heb de hele week vroege /A?/ienst" of in combinatie met een video van een gearticuleerde 'd' de waarneming van een 'd' op. De context vertelt de luisteraar dus als het ware hoe het conflict tussen de auditieve input en de relevante lexicale- en/of liplees-context opgelost dient te worden om tot een correcte waarneming te komen.

Echter, herhaalde blootstelling aan een dergelijk conflict zorgt voor een tijdelijke verschuiving in het auditieve systeem waardoor het ambigue geluid niet meer als ambigue zal worden waargenomen, zelfs als de context is verdwenen. Herhaalde blootstelling aan een geluid tussen /p/ en /t/ in het woord 'bioscoo/A?/ leidt er dus toe dat hetzelfde geluid zonder context (bijvoorbeeld in het niet bestaande woord 'dikasoo/A?/') *nog steeds* als een 'p' zal worden waargenomen. Wanneer een luisteraar het ambigue geluid tussen 'b' en 'd' in het pseudowoord 'a/A?/a' dus herhaaldelijk te horen krijgt in combinatie met een video van een spreker die /aba/ articuleert, zal de auditieve /A?/ nog steeds als /aba/ worden gehoord als de video is verdwenen. Dit effect wordt 'recalibratie' genoemd ('recalibration' in het Engels) aangezien het auditieve systeem als het ware opnieuw wordt geijkt op basis van de relevante context. De auditieve effecten zijn dus als het ware een nawerking ('aftereffect' in het Engels, in het vervolg aangeduid als na-effect) van eerdere blootstelling aan een conflict.

Hoewel na-effecten van blootstelling aan audiovisuele (in het vervolg aangeduid als 'AV') spraak enige tientallen jaren geleden al werden onderzocht, zijn

recalibratie na-effecten pas in 2003 voor het eerst gevonden (Bertelson et al., 2003). Eerdere literatuur richtte zich vooral op een ander na-effect, namelijk selectieve adaptatie. Selectieve adaptatie vindt zijn oorsprong in onderzoeken die zich op auditieve spraak richtten en wordt gedreven door een herhaling van een duidelijke spraakklank (bijvoorbeeld een /d/) waardoor /A?/ later als /b/ wordt gehoord. Mogelijk zorgt de herhaalde blootstelling aan /d/ voor een 'vermoeidheid' van de neurale netwerken die betrokken zijn bij de verwerking van de /d/-klank waardoor de /A?/ later als een als een /b/ wordt waargenomen (Eimas & Corbit, 1973). Anderen hebben daarentegen beargumenteerd dat selectieve adaptatie mogelijk een relatief simpele criterium verschuiving weerspiegelt (Diehl, 1981; Diehl et al., 1978; Diehl et al., 1980).

In tegenstelling tot recalibratie, is selectieve adaptatie niet afhankelijk van een AV conflict maar in plaats daarvan wordt het gehele effect bepaald door het auditieve segment van de stimulus zoals bevestigd in onderzoek naar selectieve adaptatie in AV spraak (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994; Shigeno, 2002).

De eerste studie naar audiovisuele spraak recalibratie

Bertelson et al. (2003) hebben als eerste recalibratie na-effecten van blootstelling aan AV spraak onderzocht. De auteurs creëerden een synthetisch spraak continuüm (9 stimuli) tussen /aba/ en /ada/ in en plaatsten de middelste stimulus (/A?/) in een videofragment van een spreker die ofwel /aba/ ofwel /ada/ uitsprak. Deelnemers aan het onderzoek werden blootgesteld aan acht herhalingen van een audiovisuele adapter (8 x /A?/V/aba/ of 8 x /A?/V/ada/) en vervolgens werden ze gevraagd om drie middelste ambigue klanken van het continuüm te identificeren als /aba/ of /ada/ *zonder* dat hier de video bij werd afgespeeld. Alle drie de stimuli (/A?/, /A?/-1 die iets meer naar /aba/ neigt en /A?/+1 die iets meer naar /ada/ neigt) werden hierbij twee keer aangeboden. Deze 'AV blootstelling – auditieve test' procedure werd vaak herhaald en proefpersonen werden tijdens het experiment dus blootgesteld aan zowel de /aba/ als /ada/ video's in combinatie met /A?/. Tijdens de auditieve test gaven proefpersonen *meer* 'b'-responsies, na blootstelling aan /A?/V/aba/ en *minder* 'b'-responsies (dus meer 'd'-responsies), na blootstelling aan /A?/V/ada/; een overduidelijk teken dat de visuele informatie de interpretatie van het geluid had beïnvloedt.

In een cruciaal controle experiment werden vervolgens ook AV congruente adapters aangeboden (A/aba/V/aba/ and A/ada/V/ada/). Aangezien er geen AV conflict in deze stimuli aanwezig is, werd selectieve adaptatie gevonden dat gedreven werd door de auditieve input. Deze adapters leidden dus tot *minder* 'b'-responsies, na blootstelling aan A/aba/V/aba/ en *meer* 'b'-responsies (dus minder 'd'-responsies), na blootstelling aan A/ada/V/ada/.

Belangrijk hierbij is dat proefpersonen de adapters zelf niet konden onderscheiden; het geluid van zowel /A?/V/aba/ als A/aba/V/aba/ werd als /aba/ waargenomen en het geluid van zowel /A?/V/ada/ als A/ada/V/ada/ werd als /ada/ gehoord. Dit is een cruciaal gegeven aangezien de auteurs nu konden concluderen dat de *tegengestelde richting* van recalibratie enerzijds en selectieve adaptatie anderzijds niet kon worden toegeschreven aan een bepaalde respons-strategie van de proefpersonen. Deze hadden namelijk niet in de gaten of ze blootgesteld werden aan een ambigue of congruente adapter.

Andere verschillen tussen recalibratie en selectieve adaptatie

Naast het feit dat recalibratie- en selectieve adaptatie na-effecten in tegengestelde richting verlopen, zijn er meerdere cruciale verschillen tussen de twee fenomenen. Ten eerste manifesteert recalibratie zichzelf al volledig na acht blootstellingen aan een adapter en wordt het effect kleiner naarmate het aantal adapters toeneemt. Selectieve adaptatie effecten daarentegen, worden groter naarmate het aantal adapters toeneemt (Vroomen et al., 2007). Een tweede verschil tussen de twee effecten werd duidelijk toen het aantal adapters gelijk werd gehouden en werd nagegaan hoe lang de na-effecten nog meetbaar waren. Recalibratie effecten verdwijnen na slechts zes auditieve test stimuli terwijl er nog steeds sprake was van selectieve adaptatie na 60 test stimuli (Vroomen et al., 2004).

Een derde verschil tussen de twee fenomenen staat beschreven in **hoofdstuk 3**. Hier werd, in plaats van synthetische natuurlijke spraak, gebruik gemaakt van zogenaamde 'Sine-wave speech' (SWS). SWS is een spraakstimulus waarin de natuurlijke akoestische rijkdom van het signaal sterk wordt gereduceerd en *alleen* de absolute basis van het geluid wordt behouden. Dit resulteert in kunstmatige stimuli die door een luisteraar niet als spraak worden waargenomen maar als computergeluiden of piepjes *tenzij* de luisteraar weet dat de stimuli in wezen spraak materiaal zijn. In dit geval hoort de luisteraar de klanken inderdaad als spraak en is dan ook niet meer in staat om de geluiden niet meer als spraak te horen (Remez et al., 1981). Het gebruik van SWS heeft dus als voordeel dat exact dezelfde stimuli kunnen worden gebruikt om spraakwaarneming te vergelijken met het waarnemen van non-spraak geluiden. In 2005 vonden Tuomainen et al. dat AV integratie van SWS en de corresponderende video's van een spreker alleen maar optreedt als de luisteraars weten dat de geluiden van spraak afkomstig zijn (de spraak groep). Luisteraars in de non-spraak groep gebruiken dus de liplees-informatie uit de video's niet wanneer ze gevraagd worden het geluid te identificeren (Tuomainen et al., 2005). Het experiment in **hoofdstuk 3** is opgezet om te bepalen of recalibratie na-effecten met SWS kunnen worden gevonden en dit bleek

inderdaad het geval. Recalibratie werd echter alleen in de spraak groep gevonden waardoor geconcludeerd kon worden dat, naar alle waarschijnlijkheid, een fonetische binding tussen de auditieve en visuele input noodzakelijk is alvorens het auditieve systeem gerecalibreerd kan worden door middel van de liplees-informatie. Voor selectieve adaptatie effecten gold echter dat ze vergelijkbaar waren in de spraak- en niet-spraak groepen waarmee werd aangetoond dat de interpretatie van een geluid als spraak/niet-spraak niet noodzakelijk is voor auditief gedreven selectieve adaptatie.

Hoe stabiel is recalibratie?

Zoals eerder aangegeven kan de lexicale context, net als de liplees-informatie, de auditieve waarneming sturen. Zo is ook onderzocht of herhaalde blootstelling aan /A?/ in combinatie met lexicale informatie in staat is om het auditieve systeem te recalibreren net als het geval is met liplees-informatie. Om dit te onderzoeken wordt /A?/ in een lexicale context geplaatst (bijvoorbeeld een klank tussen /s/ en /f/ in de context van het woord 'witlo/A?/'), herhaald aangeboden en vervolgens wordt weer getest of de waarneming van /A?/ is verschoven naar de /f/. Inmiddels is gebleken dat lexicale context inderdaad ook recalibratie kan veroorzaken (Eisner & McQueen, 2006; Jesse & McQueen, 2011; Kraljic & Samuel, 2005, 2006; Norris et al., 2003). Er is echter een belangrijk verschil tussen lexicale- en liplees-recalibratie, namelijk de gerapporteerde duur van het effect. Zoals vermeld is liplees-recalibratie van korte duur (Vroomen et al., 2004) terwijl na-effecten van lexicale recalibratie nog gevonden zijn na een interval van 25 minuten (Kraljic & Samuel, 2005). Een serie experimenten waarin lexicale- en liplees-recalibratie direct met elkaar werden vergeleken wees echter uit dat beide effecten even snel verdwenen (van Linden & Vroomen, 2007). Deze tegenstrijdige resultaten kunnen wellicht worden verklaard door een verschil in experimentele procedure aangezien de studies die langdurige lexicale recalibratie effecten rapporteerden niet alleen maar de ambigue klank gebruikten maar hun proefpersonen ook blootstelden aan de *tegenovergestelde* spraak categorie. Zo werden proefpersonen niet alleen maar blootgesteld aan de 'witlo/A?/' stimulus (die de waarneming van /A?/ naar /f/ stuurt), maar ook aan een duidelijke /s/ in het woord "radijs". De aanwezigheid van deze contrast stimulus kan inderdaad het recalibratie-effect vergroten (van Linden & Vroomen, 2007) omdat deze stimulus vergelijkingsmateriaal oplevert waartegen de kritieke /A?/ stimulus wordt afgezet. Aangezien /A?/ geen goede /s/ is (proefpersonen horen namelijk wel een goede /s/ in het woord "radijs") zal deze dus meer als /f/ worden waargenomen. Echter, een verschil in grootte van het effect wil niet zeggen dat recalibratie ook daadwerkelijk voor een langere tijd meetbaar zou moeten zijn. In **hoofdstuk 2** wordt een experiment beschreven

waarin soortgelijke contrast stimuli werden opgenomen in het liplees-recalibratie paradigma en de stabiliteit van recalibratie door de tijd heen nogmaals onderzocht werd door proefpersonen op twee momenten na blootstelling aan de AV adapters te testen; direct na de blootstelling en nog een keer 24 uur later. Dit lange tijdsinterval was gekozen aangezien een studie met lexicale recalibratie zelfs nog effecten vond 12 uur na blootstelling (Eisner & McQueen, 2006) waarin proefpersonen hadden geslapen. Slaap zou dus mogelijk en consoliderende factor van belang kunnen zijn in de stabiliteit van het recalibratie effect. Echter, de resultaten van het experiment in **hoofdstuk 2** wezen uit dat, ondanks toevoeging van de contrast stimuli, recalibratie wederom snel na blootstelling aan de adapters verdween en 24 uur later niet meer aanwezig was. Een mogelijke verklaring voor de dissociatie in de literatuur (wat betreft de stabiliteit van recalibratie) zou gevonden kunnen worden in het feit dat studies die kort durende effecten vonden alleen gebruik maakten van de kritieke adapter stimulus terwijl studies die langdurige effecten rapporteerden ook een groot aantal onbelangrijke stimuli gebruikten die de kritieke stimuli van elkaar scheidden. Het is immers bekend dat opeengepakte kritieke stimuli zwakkere leer effecten veroorzaken dan wanneer de kritieke stimuli, qua tijd, verder uit elkaar liggen (Hintzman, 1974).

Recalibratie van visuele spraak door auditieve spraak

De eerste studie waarin visuele recalibratie van auditieve spraakwaarneming werd aangetoond (Bertelson et al., 2003) liet zich voor een groot deel leiden door de gedachtegang van eerdere bevindingen met betrekking tot het zogenaamde 'buikspreker effect'. Dit effect houdt in dat een visuele stimulus de waargenomen locatie van een tegelijkertijd aangeboden geluid kan verschuiven. Net zoals een buikspreker de illusie wekt dat het stemgeluid uit de mond van de pop komt, wekt een visuele stimulus (bijvoorbeeld een lampje) de illusie dat het geluid (een simpele toon) uit de buurt van het lampje lijkt te komen terwijl dat in werkelijkheid niet het geval is. Herhaalde blootstelling aan dit soort audiovisuele 'locatie-conflicten' levert na-effecten op. Dat wil zeggen, de waargenomen locatie van de piep verschuift in de richting van de visuele stimulus, ook wanneer deze visuele stimulus gedurende de test niet meer aanwezig is (zie bijvoorbeeld Bertelson, 1999; Bertelson et al., 2006; Radeau & Bertelson, 1974, 1976) en het was deze gedachtegang die voor een groot deel ten grondslag lag aan onderzoek naar de recalibratie effecten van visuele spraak op auditieve spraakwaarneming. Voor het buiksprekereffect bleek het omgekeerde echter ook waar te zijn; de waargenomen locatie van een visuele stimulus kan verschuiven in de richting van een toon op een andere locatie en kan ook recalibratie na-effecten induceren (Radeau & Bertelson, 1987). In **hoofdstuk 4** wordt een experiment beschreven waarin

werd nagegaan of deze omgekeerde situatie ook voor spraak geldt. In plaats van een ambigue auditieve stimulus werd nu een ambigue video tussen twee alternatieven in gebruikt. Deze video werd gecombineerd met twee duidelijke spraakgeluiden en de waargenomen liplees-identiteit van de video verschoof inderdaad in de richting van het geluid. Herhaalde blootstelling aan de AV adapters leverde ook nu na-effecten op. Met andere woorden, wanneer het geluid na herhaalde blootstelling afwezig was, werd de identiteit van de ambigue video's nog steeds beoordeeld in de richting van de identiteit van het spraakgeluid. Uit deze bevindingen kan worden geconcludeerd dat de binding tussen het auditieve en visuele spraak signaal dusdanig robuust is dat beide signalen de waarneming van het andere signaal kunnen beïnvloeden en na-effecten kunnen veroorzaken.

Recalibratie en de ontwikkeling van het brein

Een aantal studies hebben aangetoond dat baby's al snel na de geboorte in staat zijn om het auditieve en visuele spraaksignaal te integreren in één fonetische waarneming (Desjardins & Werker, 2004; Kuhl & Meltzoff, 1982; Rosenblum et al., 1997). Baby's van vier maanden oud, die tegelijkertijd twee video's te zien krijgen van een spreker die een klinker uitspreekt, zijn in staat om het goede gezicht aan een auditief aangeboden klinker te koppelen (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999) en zelfs baby's van twee maanden oud kunnen de correspondentie tussen auditieve en visuele spraak detecteren (Patterson & Werker, 2003). Het is echter ook bekend dat de impact die liplees-informatie heeft op de auditieve spraak verwerking groter wordt naarmate baby's opgroeien (Massaro, 1984; McGurk & MacDonald, 1976). Zo bleek bijvoorbeeld dat 8-jarige kinderen auditieve spraak recalibratie vertonen maar 5-jarigen niet (van Linden & Vroomen, 2008).

Er zijn aanwijzingen dat minder goede lipleesvaardigheden gerelateerd zouden kunnen zijn aan een lagere leesvaardigheid (de Gelder & Vroomen, 1998). Ook is het bekend dat fonetische spraak categorieën minder goed gedefinieerd zijn bij mensen met dyslexie, dat gekenmerkt wordt door een lagere leesvaardigheid (zie bijvoorbeeld Bogliotti et al., 2008; de Gelder & Vroomen, 1998; Godfrey et al., 1981; Vandermosten et al., 2010; Werker & Tees, 1987). Problemen met liplezen zouden dus tot gevolg kunnen hebben dat spraakgerelateerde noodzakelijke aanpassingen in het auditieve systeem niet worden gemaakt. Dit zou dan wellicht resulteren in de minder goede definitie van spraak categorieën, zoals bij mensen met dyslexie het geval is. Echter, het zou ook kunnen zijn dat de minder goed gedefinieerde spraak categorieën ten grondslag liggen aan de liplees-problemen aangezien dyslexie gerelateerde problemen met spraakwaarneming al vlak na de geboorte aanwezig zijn (Guttorm et al., 2003;

Leppanen et al., 1999) en dus mogelijk de ontwikkeling van andere aspecten die met spraak te maken hebben hinderen (Guttorm et al., 2005). Naar aanleiding van deze gedachtegang werd in het experiment in **hoofdstuk 6** het mogelijke verband tussen dyslexie en recalibratie onderzocht. Een auditieve identificatie taak van het /aba/-/ada/ continuüm wees uit dat de /b/-/d/ spraakcategorieën inderdaad minder goed gedefinieerd waren in de dyslectische groep vergeleken met een groep vloeiende lezers zoals eerder al aangetoond (zie bijvoorbeeld de Gelder & Vroomen, 1998; Godfrey et al., 1981; Werker & Tees, 1987). Echter, recalibratie effecten en visuele identificatie van de video's was vergelijkbaar in de twee groepen. Hoewel er dus geen aanwijzingen werden gevonden dat liplees-vaardigheden aangetast zijn in mensen met dyslexie, is het belangrijk om in acht te nemen dat eerdere studies die wel dyslexie gerelateerde problemen met liplezen rapporteerden een moeilijkere taak gebruikten dan een visuele identificatie taak (Mohammed et al., 2006) of een groep kinderen testten in plaats van volwassenen (de Gelder & Vroomen, 1998).

Neurale mechanismen

AV spraak integratie is extensief onderzocht door gebruik te maken van meetmethoden die hersenactiviteit in kaart kunnen brengen. Zo is gevonden dat de liplees-context de verwerking van het auditieve signaal in de auditieve cortex al heel snel na het begin van het geluid (na ±100 milliseconden) kan moduleren (Besle et al., 2004; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). Er zijn echter steeds meer aanwijzingen dat deze vroege modulatie wellicht niets te maken heeft met de fonetische integratie van de signalen maar in plaats daarvan een lagere orde effect van visuele voorspelling reflecteert. Liplees informatie komt namelijk altijd eerder dan het spraakgeluid aangezien de spreker eerst de benodigde articulaties moet maken voordat de luchtstroom op een juiste manier wordt gemanipuleerd en het goede spraakgeluid wordt gevormd. De visuele informatie waarschuwt dus als het ware de luisteraar dat hij een geluid kan verwachten (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010). Dit lijkt inderdaad te worden bevestigd door het experiment in **hoofdstuk 7**. In dit onderzoek werd SWS aangeboden aan proefpersonen in zowel een spraak- als niet-spraak groep terwijl hersenactiviteit via EEG gemeten werd. EEG is een techniek waarbij fluctuaties in neuronale activiteit worden gemeten door elektroden op de scalp te plaatsen. In de cognitieve wetenschappen wordt EEG meestal gebruikt om zogenaamde ERP's ('Event Realted Potentials' in het Engels) te meten. ERP's zijn de gemiddelde veranderingen in het signaal die veroorzaakt worden door de verwerking van complexe stimuli. Uit de resultaten bleek dat de verwerking van het spraakgeluid na ongeveer 100 milliseconden in beide groepen proefpersonen werd

beïnvloed door de visuele informatie. Aangezien de niet-spraak groep niet wist dat de stimuli spraak waren kan dus niet worden geconcludeerd dat dit effect te maken heeft met spraakverwerking, zoals eerder ook al werd beargumenteerd (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010). Echter, ongeveer 200 milliseconden na aanvang van geluid werd er wel een verschil tussen de twee groepen gevonden. In de spraak groep werd wederom gevonden dat het visuele signaal de verwerking van het spraakgeluid beïnvloedde terwijl dit niet het geval was in de niet-spraak groep. Wellicht reflecteert dit effect dus een betere benadering van het tijdstip waarop spraakspecifieke visuele modulaties in het auditieve signaal optreden.

Op het gebied van spraak recalibratie zijn er echter nog maar een klein aantal studies gedaan die hebben getracht te achterhalen wat de betrokken hersengebieden zijn. Eén van deze studies (van Linden & Vroomen, 2007) liet in een EEG paradigma zien dat lexicale recalibratie van het auditieve signaal waarschijnlijk al in vroege perceptuele processen zijn oorsprong vindt. Een tweede onderzoekslijn gebruikt fMRI ('functional Magnetic Resonance Imaging' in het Engels) om hersengebieden aan te duiden die betrokken zijn bij audiovisuele recalibratie. De onderzoekers vonden dat bepaalde hersenactiviteit tijdens de blootsteling aan de AV adapters, waaronder activiteit in de 'inferior parietale cortex', betrouwbaar kon voorspellen of de proefpersonen de ambigue stimuli als /aba/ of /ada/ zouden waarnemen in de auditieve test die nog moest volgen (Kilian-Hütten, Vroomen, et al., 2011).

Aangezien de parietale cortex ook betrokken is bij het werkgeheugen (Jonides et al., 1998), lijkt het aannemelijk dat recalibratie (voor een deel) afhankelijk is van het werkgeheugen van de luisteraar. Deze gedachtegang komt met name voort uit het gegeven dat recalibratie in principe een kortdurend leer-effect is; de liplees-informatie 'leert' namelijk aan het auditieve systeem wat de klank zou moeten zijn. In **hoofdstuk 5** werd recalibratie onderzocht terwijl proefpersonen tegelijkertijd een extra taak deden waarbij het werkgeheugen werd belast. Als recalibratie inderdaad afhankelijk zou zijn van het werkgeheugen zou het recalibratie proces verstoord kunnen worden wanneer een tweede taak de benodigde capaciteit in het werkgeheugen opeist. De secundaire taak bestond of uit het onthouden van letters zodat het verbale werkgeheugen werd belast, of uit het onthouden van een traject dat door een stip op het scherm werd afgelegd zodat het visuo-spatiële werkgeheugen werd belast. De moeilijkheid van de geheugentaken werd opgevoerd in drie groepen van nieuwe proefpersonen en elke proefpersoon kreeg ook de standaard recalibratie taak zodat deze als vergelijking zou kunnen dienen. Uit de resultaten bleek dat de derde groep proefpersonen inderdaad meer moeite had met de geheugentaak dan de tweede groep, en dat de taken in de tweede groep ook moeilijker waren dan in de eerste groep. Ondanks de toegenomen moeilijkheidsgraad van de

geheugentaak waren recalibratie (en selectieve adaptatie) vergelijkbaar in alle groepen en waren na-effecten zonder extra taak even groot als na-effecten terwijl proefpersonen de extra taken deden. Voor selectieve adaptatie was al bekend dat een secundaire taak die aandacht vereist geen invloed heeft op de na-effecten (Samuel & Kat, 1998) en de resultaten met betrekking tot recalibratie lijken hetzelfde aan te geven. Hoewel in **hoofdstuk 5** de rol van aandacht niet op een directe manier werd getest, zijn de onaangetaste recalibratie effecten in overeenstemming met de argumenten dat AV integratie van spraak automatisch geschiedt (zie bijvoorbeeld McGurk & MacDonald, 1976; Näätänen et al., 1993; Soto-Faraco et al., 2004).

Conclusies

De hoofdstukken die in dit proefschrift staan beschreven maken, in combinatie met de eerdere literatuur, duidelijk dat recalibratie van spraaksignalen niet tot stand komt door toevalligheden in een experiment of een bepaalde groep proefpersonen. De na-effecten worden immers in elk relevant hoofdstuk gevonden ondanks dat gebruik gemaakt is van andere experimentele procedures en andere proefpersonen deelnamen aan de experimenten.

Recalibratie is een tijdelijk effect dat door de relevante spraak context wordt geïnduceerd en na-effecten zijn van korte duur (**hoofdstuk 2**). Recalibratie blijkt bi-directioneel van aard te zijn aangezien het mogelijk blijkt om de waargenomen liplees-identiteit te verschuiven naar eerder aangeboden spraakgeluiden (**hoofdstuk 4**). Ondanks dat recalibratie in wezen een leer-effect is, lijkt het werkgeheugen geen bepalende rol te spelen (**hoofdstuk 5**) en het leesvermogen van de luisteraars, hoewel dit samen lijkt te hangen met lip-lees vermogen, lijkt ook niet van invloed (**hoofdstuk 6**).

Echter, recalibratie lijkt wel afhankelijk te zijn van de fonetische binding tussen het auditieve en visuele signaal (**hoofdstuk 3**), die mogelijk ongeveer 200 milliseconden na de start van het geluid plaatsvindt (**hoofdstuk 7**).

Dankwoord
Acknowledgments

Deze thesis was niet tot stand gekomen zonder het inzicht, de begeleiding en het vertrouwen van mijn promotor, Jean Vroomen. Jean, we hebben een leuk en leerzaam traject doorlopen en ik hoop dat we in de toekomst nog iets voor elkaar kunnen betekenen. Jeroen, Mirjam en Liselotte, ook van jullie heb ik veel geleerd en we hebben erg veel lol gehad met ons 'neuro-clubje'. Ik hoop van harte dat we contact blijven houden en wie weet ligt er in de toekomst nog wel een gezamenlijk project in het verschiet. Sabine, jouw begeleiding tijdens mijn master-thesis is een geweldige opzet gebleken voor de afgelopen vier jaar.

Alle medewerkers van het departement Medische Psychologie en Neuropsychologie die de afgelopen jaren zijn gebleven, gekomen en gegaan; dankjewel voor de leuke gesprekken, wandelingen, lunches en de meer dan fijne verstandhouding. Ik denk dat de Universiteit van Tilburg blij mag zijn met zo'n departement.

Heather, thanks for supervising me at Haskins Laboratories.

Laurien, zonder jouw optimisme en geweldige steun was deze thesis er nooit geweest. Dankjewel lieverd!

Bram, Lucinda, Arno, Ilse, Tom en Ankie; jullie maken het Tilburgse leven leuk en bijzonder gezellig. Dankjewel daarvoor!

Stefan, Quinten, Lennart, Bart, Bram, Eelco, Koen en Gijs, bedankt voor de broodnodige muzikale en persoonlijke afleiding.

Pap, mam, Robbert en lieve familie, vanzelfsprekend wil ik jullie ook bedanken voor jullie steun door de jaren heen.