Recalibration of auditory phoneme perception by lipread and lexical information

Sabine van Linden

ISBN/EAN: 978-90-5335-122-2

Recalibration of auditory phoneme perception by lipread and lexical information

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg,
op gezag van de rector magnificus, prof. dr. F. A. van der Duyn Schouten,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op vrijdag 22 juni 2007
om 14.15 uur

door

Sabine van Linden
geboren op 7 mei 1978 te Tilburg

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

Commissie
Prof. dr. P. Bertelson
Prof. dr. R. Campbell
Dr. J. McQueen
Dr. J. Tuomainen

Contents

Chapter 1

General introduction.

Introduction

The ability to understand and produce speech is a unique and one of the most complex human abilities. One of the major obstacles in the perception of speech is the high level of variability of the incoming auditory speech signal. Due to, for example, size and shape of the vocal tract, flexibility of the tongue and state of the vocal folds, different speakers will produce different acoustic signals when uttering the same word or phoneme (Ladefoged & Broadbent, 1957; Peterson & Barney, 1952). In addition, speaking rate (Liberman, Delattre Gerstman & Cooper, 1956; Miller & Liberman, 1979; Summerfield, 1975), and differences in dialect or foreign accent influence the speech acoustics considerably. In most everyday settings the auditory signal is furthermore disturbed due to additional background noise. This raises the question how our perceptual system handles the acoustic variability in the speech signal so effectively.

Numerous studies have demonstrated that in order to correctly interpret degraded or ambiguous speech sounds, listeners make use of lipread or lexical information (Ganong, 1980; Sumby & Pollack, 1954). The unusual sounding speech is then perceived in congruence with visual articulatory gestures or with the lexical context information. For example, an ambiguous phoneme between /b/ and /p/ will be perceived as a /b/ when presented before 'oat', creating the word 'boat', but the same sound will be perceived as a /p/ when its presented before 'ope', since it then completes the word 'pope' (Ganong, 1980).

Only recently, it has been demonstrated that the brain also uses these sources of information for future interpretation of the ambiguous speech sound. Phoneme representations are flexibly adjusted in congruence with available lipread or lexical information in order to interpret new or unusual sounding phonemes. This process, which is referred to as recalibration, probably reflects the mechanism with which the brain deals with the endless acoustic variability in speech sounds.

The current thesis aims to expand the knowledge regarding lipread- and lexically-driven recalibration in auditory speech perception. The first part of the thesis is concerned with the fundamental properties of lipread-driven recalibration. Important issues are for example the speed at which such learning effects occur and how long they persist in time. The second part of the thesis investigates and compares lipread- and lexically-driven recalibration and their effects on online phoneme perception. As lipread and lexical information are inherently different in nature, they might affect auditory speech processing in different ways. The first chapter provides an overview of lipread and lexical effects on speech perception, perceptual learning and the aftereffects in speech perception. It furthermore provides a description of the neurological mechanisms of audiovisual speech integration and of two contrasting theories explaining lexical effects in speech perception.

Direct bias in speech perception

Perceiving speech is not solely an auditory skill. Especially under conditions of degraded or ambiguous acoustic speech input, lipread and lexical information are used to disambiguate the signal. Perception of the ambiguous speech sound is then biased towards the lipread or lexical context information, an effect referred to as 'direct bias' in speech perception.

Visual effects on online auditory speech perception

Speech is the example 'par excellence' that human information processing is multisensory. In a normal face-to-face conversation, a listener receives two sources of speech input: the acoustic signal, and the articulatory movements of the speaker i.e. visual speech. Perceiving the articulatory gestures enhances auditory speech perception (Callan et al., 2003), especially under conditions of degraded auditory input (Sumby & Pollack, 1954). Sumby and Pollack (1954) systematically varied the signal-to-noise ratio when presenting auditory-only or audiovisual speech. Word recognition was substantially better in the audiovisual speech conditions, with improvement reaching an equivalent of a 15 dB increase in the signal–to-noise ratio. Visual speech information not only affects speech perception under conditions of degraded auditory input: the influence of lipread information is so strong that it can even alter the perception of auditory clear speech tokens (McGurk & MacDonald, 1976). In their now classic study, McGurk and MacDonald, (1976), presented participants incongruent auditory and lipread syllables which resulted in fused or combined precepts' of the auditory and visual speech information. For example, when dubbing auditory /ba/ onto a face articulating /ga/, participants reported hearing the fused percept /da/. The reverse pairing, i.e. auditory /ga/ with visual /ba/, often induced the percept /bga/, a combination or blend of the acoustic and visual information. The McGurk effect demonstrates the effortless integration of auditory and visual speech cues to form a new, multimodal percept. It shows that we can't help but integrate visual speech into what we "hear". The research by McGurk and MacDonald, demonstrated that speech perception should not be considered only an auditory skill, with only a supportive role for visual information.

Several developmental studies suggest that speech is represented intermodally already very early in life, indicating that the utilisation of lipread gestures does not solely depend on prolonged experience with the corresponding relation. Infants in age of 18 to 20 weeks prefer to look at a face making articulatory gestures in congruence with the heard vowel /i/ and /a/, rather than to a face making incongruent speech gestures (Kuhl & Meltzoff, 1984). However, when all spectral information was eliminated from the acoustic signal, leaving the timing and amplitude information intact, the infants showed no preference for the congruent visual speech. Temporal information was thus not a sufficient

cue to induce the matching effect. This implies that infants recognize the correspondence between auditory and visually speech cues. In addition to the preference for congruent rather than incongruent speech gestures, five-month-old infants are also susceptible to the McGurk-illusion. Rosenblum, Schmuckler and Johnson (1997) demonstrated that infants generalized across speech tokens, which were acoustically different but perceptually similar due to articulatory gestures (see also Burnham, 2004). The speech tokens presented was an audiovisual congruent /va/ (AvVv) stimulus, and two audiovisual incongruent stimuli: auditory /ba/ dubbed onto visual /va/ (AbVv) and auditory /da/ dubbed onto a face articulating /va/ (AdVv). In previous study it was demonstrated that adults perceived AbVv as /va/ in 98% of the time, whereas AdVv was perceived as /da/ 88% of the time (Saldana & Rosenblum, 1993; Saldaña & Rosenblum, 1994). Infants generalized from AvVv to AbVv, but not from AvVv to AdVv, thus demonstrating their susceptibility to the McGurk-effect.

In addition to being present early in development, research on audiovisual speech perception has demonstrated that the auditory illusion produced by McGurk stimuli are highly robust (Bertelson, 1996). Naive participants are usually not aware of any discrepancy between the auditory and visually signals when confronted with a McGurk stimulus and report integrated percepts even when explicitly instructed to focus attention to only one of the modalities (Massaro, 1987). In their research report, McGurk and MacDonald (1976) stated that although "the analyses statistically confirmed the visual influence on auditory speech perception, the data failed to testify the powerful nature of the illusions". They furthermore stated that: "We ourselves have experienced these effects on many hundreds of trials; they do not habituate over time, despite objective knowledge of the illusion involved". The occurrence of the McGurk effect is furthermore highly resistant to all sorts of distortions of natural audio-visual speech conditions. Integration of auditory and visually presented speech is not (or minimally) affected by face-voice gender-incongruence (Green et al., 1991), by spatial separation of the auditory and visual streams (Colin et al., 2001 Bertelson, Vroomen, Wiegeraad & de Gelder, 1994), viewing distance or image size (Jordan & Sergeant, 1998, 2000), or gaze direction of the observer (Pare et al., 2003). Temporal alignment of lipread and auditory speech information does influence the occurring of the McGurk, but the temporal window under which the effect occurs is relatively wide (Munhall et al., 1996; van Wassenhove et al., 2006).

A still standing issue is whether directed attention plays a role in the binding of auditory and visual cues, as has been suggested by some researchers (Alsius et al., 2005). If indeed attention would be necessary for the binding of multi-modal speech cues, directing attention away from the stimuli or depleting attentional resources by a dual task would predictably lead to a decrease in visual influenced responses on an audiovisual speech identification tasks. Such effects have indeed been reported to occur, as for

example Tiippana et al. (2004) reported that tracing a moving leaf over a speaking face reduced visually influenced identification of a McGurk stimulus. Also, Alsius, Navarra, Campbell and Soto-Faraco (2005) reported that the proportion of visually influenced responses was reduced if participants performed an unrelated visual or auditory task. Most likely, however, in both studies the additional tasks interfered with lipreading as such, which on its turn caused the decrease in visually influenced responses (Vroomen, submitted). Obviously, when lipreading is complicated by an additional task this leads to a decrease in visual influence. It is thus not yet clear, whether the additional tasks hindered integration of the speech tokens through attentional modulation or whether the effect was caused by complicating lipreading.

Visual speech provides, among others, information regarding place of articulation, a feature of speech which is visually distinctive, but which is acoustically fragile (Grant et al., 1998; Vroomen, 1992). The place of articulation feature of a phoneme specifies the point of obstruction created by the articulators, like the tongue, lips and teeth, in the vocal tract. Place of articulation, together with manner of articulation and voicing determine the specific sound of a phoneme. Whereas voicing and manner are rather robust auditory cues but difficult to see, the place feature is a distinctive visual cue. For example, lip closure as in the pronunciation of /b/, is a clearly visible gesture. The most striking effects of visual information are reported for visual bilabials (/b/ or /p/) dubbed onto auditory presented non-labial consonants like /t/ and /d/. Notably, the auditory cues for place of articulation are the first speech characteristics to get lost in noisy backgrounds; lipread and auditory cues therefore share a complementary relationship.

In natural speech, the articulatory and acoustic information do not exactly coincide in time: visual speech information often is available before the acoustic consequence of this movement (Bell-Berti & Harris, 1981; Munhall et al., 2004). Articulatory movements can thus prime and facilitate auditory speech perception. Indeed, when measuring auditory evoked potentials while presenting auditory-only and audiovisual speech, van Wassenhoven et al. (2003) reported a reduction of amplitude for the N1/P2 complex and a temporal facilitation for the P1/N1/P2 complex for audiovisual as opposed to auditory-only speech. Visual information thus affected auditory processing of speech at a very early stage, already between 50 and 100 ms post-auditory onset.

In addition to phonetic feature information, visual speech provides cues regarding the temporal properties of the acoustic signal. There is a high degree of correlation between lip and head movements and the acoustic amplitude envelope (Munhall et al., 2004; Rosen, 1992; Summerfield, 1992). For example, Munhall et al. (2004) reported a correlation between head movements and intelligibility of speech in noise. In their study, participants were asked to identify as many words as possible in a sentence. Sentences

were presented in four different conditions: an auditory-only condition and three audiovisual conditions in which a speaker made appropriate articulatory gestures. The head motion of the talker was normal, doubled in amplitude or non-existing. It was observed that the intelligibility of the speech in noise varied as a function of the head-motion. Performance in all audiovisual conditions was better than in the auditory-only condition. In addition, in the natural head movement condition, performance was better than in the no-head movement or double head-movement condition. The production data were also analyzed, showing a systematic relationship between head movements, F0, and the amplitude of the voice.

Lexical effects on online auditory speech perception

The first demonstration that stored lexical knowledge influences the perception of degraded speech was provided by Warren (1970b). Warren presented participants a sentence in which part of a speech utterance was deleted and replaced with a cough, a buzz or a tone. Participants did not perceive the replacement of the speech sound and could not localize the replacement. Even when participants were notified that a speech sound was replaced with noise, they were not able to distinguish the illusory sound from the real phoneme. The phenomenon was called phoneme restoration. Only when a silent gap replaced the speech segment, the missing sound could be localized. In a later study by Samuel (1981a) participants were not able to discriminate between words containing truly deleted and replaced phonemes and intact words when the noise was superimposed onto the particular phoneme, illustrating the perceptual pervasiveness of the illusion.

As described before, Ganong (1980) demonstrated that lexical information influences the perception of ambiguous speech sounds. In his study, participants were presented a da-ta continuum spliced onto either /sk/ or /sh/ resulting in a /dask/ -/task/ and a /dash/- /tash/ continuum. The midpoint tokens on the dask-task continuum were more often perceived as the word "task" than as the non-word "dask" and more often as the word "dash" than as the non-word tash" on the /dash/-/tash/ continuum.

Lexical effects on phoneme perception have furthermore been reported to occur in phoneme-monitoring tasks. Due to lexical activation, a target phoneme is detected faster when it is presented in a word rather than in a non-word (Rubin et al., 1976), and is also detected faster when presented in high-frequency words than in low-frequency words (Dupoux & Mehler, 1990). Also, monitoring for phonemes when the target phoneme is positioned at the end of a non-word is faster the more similar the non-words are to real words (Connine et al., 1997).

Since the initial reports of phonemic restoration, the Ganong-effect, and lexical effects in phoneme monitoring, the main question regarding these online phenomena in

speech perception is when and how top-down lexical information is integrated with the bottom-up acoustical speech input. Although most theorists agree that several levels of processing are involved in the decoding of auditory speech, there is an ongoing debate regarding the communication between these levels. There are two main views regarding this issue, with one of them stating that speech processing is an interactive process in which there is direct feedback from lexical levels onto prelexical processing stages. According to this account, the Ganong-effect reflects a change in phoneme perception due to lexical feedback on the phonemic level (Ganong, 1980; McClelland & Elman, 1986; McClelland et al., 2006; Samuel, 1997, 2001). According to the other view, speech processing does not use online feedback from lexical stages onto earlier stages in speech processing. Speech perception is considered to be a feed-forward process, where lexical and prelexical information are integrated at a post-perceptual decision stage (Cutler et al., 1987a; Norris et al., 2000).

To investigate whether lexical information indeed affects processing on the earlier phonemic level, Samuel (1997) investigated for the occurrence of adaptation effects produced by exposure to lexically-restored phonemes. In the selective speech adaptation paradigm as used in this study, listeners identify a range of speech tokens of a particular phoneme continuum before and after exposure to an endpoint speech token of that continuum (Eimas & Corbit, 1973). Selective speech adaptation reveals itself as the reduced tendency to report the phoneme which was heard during the exposure phase on the subsequent post-test trails. Samuel (1997) presented participants with words like /alphabet/ and /armadillo/, in which case the stop-consonant /b/ or /d/ was deleted and either replaced by noise or by silence. The effect of this exposure was measured on a subsequent categorization task of a /bi/ -/di/ continuum. It was hypothesized that if lexical information affects phoneme perception at a prelexical stage, the restored phonemes in words containing the noise replacement would produce selective speech adaptation (Eimas & Corbit, 1973). As aftereffects are measured on the isolated phonemes in the absence of lexical context information, a change in perception of these phonemes could not be attributable to lexical influences on a post-perceptual decision stage. A change in the perception of a phoneme as a consequence of top-down lexical effects would therefore reflect a change in perceptual processing due to top-down feedback mechanisms. For the words containing the silence gap, no aftereffects were anticipated as no phonemic restoration occurs for such stimuli. Results confirmed these predictions: after exposure to words in which the /b/ was replaced by noise, participants gave fewer /b/ responses on the categorization task than before exposure. Exposure to words in which the /d/ was replaced by noise produced a similar response pattern: fewer /d/ responses after exposure. Selective speech adaptation was thus induced by a perceptually restored phoneme. This result was taken to imply that the lexical information activated prelexical

phonemic representations, thereby inducing selective speech adaptation. Samuel argued that such effect would only be possible when there is lexical feedback onto prelexical processing stages, and that the results were therefore supportive of an interactive view of speech perception.

Recalibration by pairing

In addition to our unique language ability, humans are gifted with an impressive ability to learn and to adapt to the changing demands of the surrounding environment. There are many different kinds of learning which are generally divided into explicit and implicit learning. Explicit learning involves the formation of declarative memories regarding events, objects or people and requires awareness. Implicit learning on the other hand, does not require the involvement of conscious awareness, does not lead to declarative memory of the learned entry, but leads to improvements or changes in sensory perception (Gibson, 1963; Gilbert et al., 2001). One type of implicit learning is perceptual learning. In perceptual learning, one's ability to discriminate between particular features of a stimulus is improved with training. Training here means practice with or repeated exposure to a particular stimulus. As for speech perception, training improves discrimination accuracy on novel, non-native phonetic-contrasts, demonstrating that adults are still able to acquire new phoneme contrasts. As an example, Logan, Lively and Pisoni (1991) trained native Japanese listeners to identify the /r/ - /l/ contrast with a two-forced choice identification task. Prolonged training improved discrimination ability considerably on the initially difficult task. The improvement of discrimination on a non-native phoneme contrast has been replicated several times and the usual procedure involves a long training regime. The training phase in the studies by Logan et al. (1991) and Bradlow et. al (1997) for example all consisted of many sessions spread over a period of 3 to 4 weeks. Still, although discrimination is improved considerably, identification accuracy is still substantially poorer than achieved by native speakers (Bradlow *et al.*, 1997).

In addition to improved discrimination ability on a new phoneme contrast, other studies have demonstrated that with training, listeners adapt to synthetic speech, (Greenspan et al., 1988; Sheffert et al., 2002) compressed speech, talker- and speaking rate changes (Dupoux & Green, 1997), and foreign-accented speech (Bradlow et al., 1997; Clarke, 2000). All these studies typically expose participants to unusual sounding speech after which improvements in intelligibility or discrimination are observed.

Another form of implicit learning is recalibration by pairing (Epstein, 1975; Wallach & Karsh, 1963; Wallach et al., 1963). Recalibration by pairing emerges as a consequence of exposure to conflicting information inputs regarding a stimulus characteristic. Recalibration was first reported by Wallach and Karsh (1963) who

investigated perceptual learning for depth-cues using a pretest- exposure - posttest paradigm. During pretests, participants made depth-judgments based on a single depth cue. In the exposure phase, or the pairing phase as it was referred to by the authors, the original depth cue was paired with a second cue which provided discrepant information regarding the distal property of the stimulus. After a period of exposure to this conflict situation, participants were again presented with the first cue only and were asked to make depth-judgments. Comparing pretest-and post-test judgments revealed that the posttest judgments were shifted towards the direction of the paired cue presented during the exposure phase.

In addition to recalibration by pairing within a single modality, recalibration by pairing also occurs across modalities (Epstein, 1975; Radeau & Bertelson, 1974; Wallach, 1968). Under conditions of discrepant information provided by two or more modalities regarding a specific characteristic of a stimulus (i.e. identity, location, movement etc.), the information processing in the modalities involved can be recalibrated. An example is the so-called ventriloquist-aftereffect in which discrepant visual information regarding the perceived location of a sound source recalibrates auditory sound localization (Frissen, 2005; Radeau & Bertelson, 1974; Recanzone, 1998). In a ventriloquist situation, a sound and light flash are presented simultaneously, but at different spatial locations. The perceived location of the sound is than shifted towards the direction of the light flash (Bertelson & Radeau, 1981). Exposure to this conflict situation produces a shift in the perceived location of a sound which is presented in isolation on a subsequent localisation task. The perceived location of the sound is then shifted towards the previous location of the light-flash.

Recalibration has been demonstrated to occur for many perceptual functions (see Bedford, 1999 for a review). Since the observed perceptual shifts are compensatory such that the perceived discrepancy is reduced, recalibration probably reflects a mechanism by which intermodal relationships are kept in correct coordination with each other. Conflict situations are not only created in the laboratory for the purpose of studying the process of recalibration: errors in coordination between or within sensory modalities may arise naturally due to for example growth of the body, sensory handicap, or when the information provided by one of the modalities or about a relevant cue is indistinct. (Bedford, 1999; de Gelder & Bertelson, 2004; Held, 1965). Aftereffects induced by exposure to conflicting information are furthermore taken as a strong indication of the perceptual nature of crossmodal phenomena. Since the effects are measured on unimodal posttests, they cannot be attributed to post-perceptual decisions or judgemental processes (Bertelson, 1996).

Recalibration in speech perception

The occurrence of recalibration occupies an important place in the current picture of cross-modal interactions. On the one hand, it reveals the existence of a genuine perceptual contribution to the causation of such interactions. This argument has been used (Bertelson, 1999; Bertelson & Radeau, 1976; Radeau, 1994a; Vroomen, 1999) against reductionist interpretations of cross-modal biases in terms of post-perceptual processes (Choe, Welch, Gilford, & Juola, 1975; Welch, 1999; Welch & Warren, 1980). On the other hand, recalibration probably plays a central role in developing and later maintaining cross-modal co-ordination (Held, 1965). Recalibration thus links cross-modality research with evolutionary accounts of perceptual function.

Visual recalibration of auditory phoneme perception

Recalibration occurs when discrepant cues or modalities are paired. In speech perception, listeners have the acoustic signal, lipread speech and lexical knowledge at their disposal. It has recently been demonstrated that exposure to a discrepancy between these sources of information causes recalibration in phoneme perception. (Bertelson et al., 2003; Kraljic & Samuel, 2005; Norris et al., 2003).

The study by Bertelson et al. (2003) was the first study to report visually induced recalibration in the perception of speech sounds. In order to isolate effects of recalibration, an auditory ambiguous phoneme was combined with lipread information. Exposure to such audiovisual ambiguous phonemes resulted in an increase in visual consistent identification responses on auditory-only post-test. Lipread information thus affected how to interpret the ambiguous phoneme when it was later presented alone.

In their study, Bertelson et al. (2003) exposed participants to an auditory ambiguous phoneme intermediate between /aba/ and /ada/ (henceforth an auditory ambiguous phoneme will be represented as /A?/) which was dubbed onto the video of a face articulating either /aba/ or /ada/ (resulting in the audiovisual ambiguous adapters A?Vb and A?Vd). A short period of exposure (eight trails) to these audiovisual ambiguous adapters produced positive aftereffects in subsequent auditory-only post-tests. That is, after exposure to A?Vb, the auditory ambiguous testtokens were more likely to be judged as /aba/ than after exposure to A?Vd. Exposure to the audiovisual ambiguous exposure stimuli thus led to an increase of adapter consistent responses on auditory-only posttests. In a second experiment, participants were also exposed to audiovisual congruent adapters. Auditory clear /aba/ was dubbed onto a face articulating /aba/ (AbVb) and auditory clear /ada/ was dubbed onto the video of a face articulating /ada/ (AdVd). Exposure to these adapters produced negative aftereffects: i.e., less adapter-consistent responses on the posttests, thus revealing selective speech adaptation. Selective speech adaptation is an aftereffect that depends upon the repeated presentation of an auditory clear speech

sound, (Eimas & Corbit, 1973). Importantly, participants were not able to discriminate between the audiovisual congruent and audiovisual ambiguous adapters in a subsequent stimulus identification and discrimination task. Practically all A?Vb and AbVb were judged as /aba/, 98% and 100% respectively, and all A?Vd and AdVd stimuli were judged as /ada/, (both 100% /ada/ response). Even more convincing, in a subsequent ABX-task participants had to discriminate between A?Vb and AbVb an between A?Vd and AdVd. Discrimination performance was near chance level, 52% correct, implying the participants did not perceive any difference between these stimuli. Importantly, although audiovisual ambiguous and audiovisual congruent adapters were perceived similarly, they produced aftereffects in opposite directions. Post-perceptual response strategies could therefore not be held responsible for the observed aftereffects since participants were not even able to discriminate between the audiovisual ambiguous and congruent exposure stimuli.

Lexical recalibration of auditory phoneme perception

In addition to visually-driven recalibration, lexically-driven recalibration of auditory speech perception has recently been demonstrated (Norris et al., 2003; Kraljic & Samuel, 2005; Eisner, 2006; Eisner & McQueen, 2005, 2006). In the initial report, (Norris et al., 2003), the final fricative of 20 critical words had been replaced by an ambiguous sound intermediate between /f/ and /s/. One group of listeners heard ambiguous /f/-final words (e.g., /witlo?/, from witlof, chicory) and unambiguous /s/-final words (e.g., naaldbos, pine forest). Another group heard the reverse (e.g., ambiguous /naaldbo?/, unambiguous witlof). Listeners who had heard /A?/ in /f/-final words were subsequently more likely to categorize ambiguous sounds on an /f/-/s/ continuum as /f/ than those who heard /A?/ in /s/-final words. This was the first study to indicate that listeners made use of lexical information to adjust phoneme representations. The ambiguous fricatives were also spliced after non-words. Listeners exposed to these non-words showed no shift in perception of the continuum. This observation indicated that the observed effect indeed depended upon the lexical context.

Recent results indicate that the lexically-guided perceptual adjustments are speaker/token specific and probably rely on the acoustic properties of the stimuli used. Lexically-driven perceptual learning on an ambiguous fricative between /s/ and /f/ produced by a female speaker did not affect the perception of speech produced by a male speaker and vice versa in the study by Eisner and McQueen (Eisner & McQueen, 2005). Kraljic and Samuel (2006) however, reported generalization of lexically-driven perceptual learning on an ambiguous stop consonant between /t/ and /d/. The perceptual learning effect induced on this ambiguous phoneme produced by a particular speaker generalized to speech produced by another speaker. Moreover, the effect also transferred to another stop-consonant phoneme contrast: perceptual learning effects were also observed on a /b/

- /p/ continuum subsequent to /t/ or /d/ training. Critically, fricatives as used in the study by Eisner and McQueen (2005) are cued by spectral information which is relatively sensitive to differences between speakers. The temporally defined voice-onset distinction of the two stop-consonants on the other hand, provides less speaker-specific information and the /b/ -/p/ stop continuum varies in voice-onset time in a similar way as the /d/ - /t/ continuum. Hence, learning regarding voice-onset on the /d/ - /t/ contrast could be applied to the /b/ - /p/ continuum and also generalized to a different voice. An acoustically-based mechanism for lexically-driven perceptual learning or recalibration in phoneme perception would predict such results. Such an acoustic mechanism would be beneficial in the purpose of recalibration which is to overcome the high level of variability in the acoustic speech signal. The system can make use of different types of information and depending on the acoustic overlap between phonemes, perceptual learning can generalize to different speakers and phonemes.

Selective speech adaptation

In addition to recalibration which reflects the adaptive and flexible nature of speech processing, there is another aftereffect, namely selective speech adaptation. The occurrence of selective speech adaptation does not serve an adaptive function as with recalibration, but it has been taken as the demonstration of the existence of specialized phonetic feature detectors sensitive to the specific features of a particular phoneme.

In selective speech adaptation it is the repeated presentation of a clear phoneme which results in a reduced tendency to report that phoneme on subsequent auditory posttests (Eimas & Corbit, 1973). Eimas and Corbit exposed listeners to the repeated presentation of one of the endpoint syllables of a /ba/ - /pa/ continuum. Participants were subsequently asked to identify all the tokens of the continuum. After the repeated exposure to /ba/, listeners identified fewer tokens of the continuum as /ba/ than when they had been exposed to /pa/. In analogy with well-known contrast effects in the visual modality, like for example colour perception (Eimas et al., 1973) the authors suggested that selective speech adaptation reflected fatigue of specialized phonetic feature detectors that become desensitized by repeated stimulation.

Selective speech adaptation also generalizes to other phoneme categories. For example, the repeated presentation of /da/ results in fewer /ba/ responses on a subsequently presented /ba/-/pa/ continuum, presumably because /da/ fatigues the voicing feature. The results are thus not specific for the phoneme of the exposure series, but may reflect a more general feature of the adaptor. The repeated presentation of the feature to which a speech detector is sensitive, then fatigues the detector and thus reduces its sensitivity. As a consequence, speech tokens are differently assigned to

phonetic categories, especially when the phoneme to-be-categorized contains information to which both detectors are sensitive. This change in categorization would thus result in an increased number of responses for the opposite category.

Nature of selective speech adaptation

As the occurrence of selective speech adaptation was taken to reveal the existence of specialized feature detectors, the question regarding the nature of these detectors emerged. Some researchers questioned the speech-specific nature of these detectors and advocated that selective speech adaptation occurs at an acoustic rather than a phonetic level. (Cooper, 1974; Roberts & Summerfield, 1981). In addition, some investigators questioned the existence of specialized feature detectors altogether, attributing the effect to response bias or contrast effects (Diehl et al., 1978). Although the research on selective speech adaptation increased our knowledge on speech perception, these issues have not been resolved completely.

According to the contrast account of selective speech adaptation, the repeatedly presented clear phoneme (the endpoint stimulus) serves as a good exemplar of its phoneme category. It then functions as a reference against which subsequent phonemes are compared. The criteria for a speech sound to be included in that category will than become more stringent. As a consequence, the ambiguous stimuli will be perceived as belonging to the opposite category. In support of this view, it was demonstrated that the presentation of a single adapter also produced selective speech adaptation, even when the adapter was presented after the to-be-labelled stimulus (Diehl, 1978). It was reasoned that a single presentation of stimulus should not produce fatigue, especially not when that stimulus is presented after the identification token. Since results were comparable to the usual selective speech adaptation procedure, explanations of fatigue were questioned.

The level at which selective speech adaptation occurs, either acoustic or phonetic, is also controversial. Although Eimas and Cooper (Eimas et al., 1973) initially suggested the effect to be phonetic, some researchers attributed adaptation effects to low-level auditory processes with no involvement of speech-specific mechanisms. Evidence favouring the acoustic account is that similar selective speech adaptation effects can be obtained by speech and non-speech sounds (Samuel, 1986; Samuel & Newport, 1979) and that the more acoustically similar the adapter and testtoken stimuli are, the bigger the magnitude of the effect (Sawusch & Jusczyk, 1981; Sawusch, 1977). Ohde and Sharf (1979) investigated whether selective speech adaptation is sensitive to low-level acoustic parameters. If so, this would be indicative of an acoustic basis for selective speech adaptation. Increasing the intensity of the adapter and increasing the number of adapter repetitions significantly increased the magnitude of the adaptation effect. Furthermore, no

correlation was found between the effect size and the percept of the adapter.

Visual effects in selective speech adaptation

To disentangle phonetic from acoustic accounts of selective speech adaptation, researchers tried to disassociate the acoustic structure from the phonetic percept of speech stimuli. In order to do this, one could either create perceptually similar but acoustically different stimuli (Sawusch, 1977; Sawusch & Jusczyk, 1981), or acoustically similar but perceptually different stimuli (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994) and measure their respective adaptation effects. If selective speech adaptation depends upon the acoustic structure of the stimuli, the size of the selective speech adaptation effect would be positively related to the acoustic similarity (i.e. degree of spectral overlap) between the adapter- and test- stimuli. Such a relation was not always found and selective speech adaptation effects have been reported for adapters that were acoustically different to the posttest tokens (Samuel, 1986; Samuel, Kat & Tartter, 1984; Sawuch & Jusczyk, 1981; Diehl, 1978).

An elegant study, conducted by Roberts & Summerfield, (1981), utilized the McGurk-effect to obtain acoustically similar but perceptual different adapters. Participants were exposed to an incongruent audiovisual pair consisting of auditory /bɛ/ dubbed onto the visual presentation of a face articulating /gɛ/, inducing the percept /dɛ/. Adaptation effects obtained with this adapter were compared to the effects obtained with audiovisual congruent /bɛ/ and /dɛ/ adapters (AbVb and AdVd adapters). Participants were also adapted to unimodal auditory speech tokens /bɛ/ and /dɛ/. Importantly, the incongruent McGurk-adapter produced similar adaptation effects as the congruent AbVb adapter implying that selective speech adaptation depends on the acoustic characteristics of the stimuli and not on the perceived phonetic identity. Furthermore, exposure to the unimodal auditory adapters resulted in the same reduction as the audiovisual congruent and incongruent adapters, whereas the visual adapters produced no effect whatsoever. These results were taken to imply that selective speech adaptation depends on the acoustic nature of the stimuli and not on the phonetic percept. This result was later replicated by Saldana and Rosenblum (1994) with different audiovisual stimuli as adapters producing stronger McGurk-effects. Despite the fact that the visual component indeed resulted in stronger effects on phoneme perception than the stimuli used by Robert & Summerflield (1981), aftereffects were as before: Selective speech adaptation depended on the auditory component of the stimuli, and not its percept.

In both studies, lipreading had no effect on aftereffects. This might at first sight be somewhat surprising, because in the McGurk stimulus there is conflict between the heard and seen information that might trigger recalibration. The incongruent audiovisual adapter could thus potentially produce both selective speech adaptation (because there is

an unambiguous sound) and recalibration (because there is a conflict). The possibility of summation of selective speech adaptation and recalibration due to exposure to a McGurk stimulus was further investigated in the present thesis.

Principles of recalibration and selective speech adaptation.

Although recalibration and selective speech adaptation are both observable as aftereffects in speech perception they probably depend on different processes (Bertelson et al., 2003). Whereas recalibration represents the adaptive adjustment of phoneme boundaries, selective speech adaptation most probably reflects fatigue of some of the relevant speech processes. The study by Bertelson et al.(2003) indeed obtained a dissociation between the two effects, just by manipulating the ambiguity of the auditory speech token. By exploring the basic properties of selective speech adaptation and recalibration in speech perception in the present thesis, we not only add to the understanding of both aftereffects but this approach also allows us to compare their underlying mechanisms.

Unity assumption

The assumption of unity states that for direct bias or recalibration to occur, the observer must be under the supposition that the conflicting input signals arise from a single event or identity (Welch, 1972; 1978). According to this theory, the perceptual binding of the signals depends upon the degree of consistency between the input signals. The more properties are shared, the more likely it is that the signals share a common source, and the more likely it will be that the signals are bound into one multimodal percept. (Bedford, 1989; Bertelson, 1999; Welch, 1999). Temporal and spatial congruency are proposed to be the two most important properties the signals share in this respect, (Radeau, 1994; Vatakis & Spence, in press), although recent research suggests that the spatial properties may not be crucial (Vroomen & Keetels, 2006). Integration of acoustic and visual speech cues as occurring in the McGurk effect are more likely to depend upon temporal proximity of the signals, while unaffected by spatial segregation (Bertelson et al., 1994; Colin *et al.*, 2001). As the assumption of unity is important for signal binding in online perception and possibly for recalibration, one might expect to find a positive correlation between the magnitude of direct bias and recalibration. Although this proposition seems intuitive, the existence of such a relation has not been investigated before. For speech perception no data on this issue is present, as recalibration for auditory speech perception has only been reported only recently (Bertelson et al., 2003; Kraljic & Samuel, 2005; Norris *et al.*, 2003). The present thesis provides some insights on this topic.

Perceptual dominance

When confronted with a bimodal conflict situation, the input provided by one modality is usually more dominant in the resulting multimodal percept than the input provided by the other modality. In the ventriloquist-aftereffect for example, visual information produces a large shift in the localization of sounds. This phenomenon is referred to as visual capture. The opposite result, a shift in visual localization induced by incongruent spatial auditory information, has also been observed (Radeau and Bertelson, 1974) but the effect of auditory capture on visual localization is much smaller. The modality appropriate hypothesis states that the relative influence of two sources of information depends upon the modality appropriateness for the perceptual task at hand. Since visual information is the more accurate or appropriate modality for spatial localisation, visual information will dominate auditory information in a localisation task, as occurs in the ventriloquist situation. In contrast, temporal resolution is higher in the auditory than visual modality. Auditory information therefore dominates visual information in tasks requiring temporal processing, both in online processing as in aftereffects. (Morein-Zamir et al., 2003; Vroomen & de Gelder, 2004b; Vroomen & Keetels, 2006).

The impact of the information provided by the different modalities is not only an inherent property of the particular modalities involved, but also depends upon the reliability of the input signal. For example, when the visual information in the ventriloquist situation is blurred, making the visual cues a less reliable estimate for localization, the visual impact over auditory information decreases and can even reverse (Alais & Burr, 2004). The observed increase of visual influence on auditory speech perception when background noise levels increase (Callan et al., 2003; Sekiyama et al., 2003b; Sumby & Pollack, 1954) is in accordance with this idea. Likewise, and as stated before, lipread information has the strongest effect on the perception of place of articulation, a visually distinct but auditory somewhat ambiguous feature. It thus seems that the brain uses each modality's specific strength, such that the information that is most accurately encoded in one modality influences perception in the other modality (see also Ernst & Banks, 2002). It is thus the combination of the modality and the reliability of the stimulus that determine the perceptual capture or dominance and not the modality or the stimulus per se. Ernst and Bülthoff (2004) suggested the use of the term "estimate precision' instead of "modality appropriateness" or "modality precision", since, as they stated, perceptual dominance is determined by the estimate and how reliable it can be derived within a specific modality for a given stimulus. It is thus not an inherent property of the modality itself.

Importantly, studies on lipread and lexical recalibration, by Bertelson et al. (2003), Norris et al. (2003) and Kraljic and Samuel (2005), changed the perception of an ambiguous speech sound. Since the auditory component is ambiguous, lipread or lexical

information is the more reliable source of information regarding phoneme identity, and thus creating the right conditions for recalibration to occur.

The role of the both the acoustic and visual component for recalibration and selective speech adaptation is further investigated in chapter 2. In the second part of this thesis, the effects on speech perception of lipread and lexical information are investigated and compared.

Time course

Another issue regarding recalibration and selective speech adaptation in speech perception is their development over time. In his review on recalibration by pairing Epstein (1975) states that the magnitude of the recalibration effect is positively related to the duration of exposure to the paired inputs. Thus, prolonging exposure should result in bigger aftereffects indicative of recalibration. The relation between exposure duration and magnitude of the aftereffect was explored in chapter 5 for both visually-driven recalibration and selective speech adaptation. Bedford (1999), furthermore stated that recalibration effects are typically acquired fast, already after a short period of exposure. Investigating the build-up course of recalibration allowed us to see how fast recalibration is acquired. For selective speech adaptation, there is some evidence that the effect becomes stronger when exposure is prolonged. In the study by Ohde and Sharf (1979) either 5, 32 or 95 adapter presentations preceded posttesting for selective speech adaptation. Although the effect increased from 5 to 32 adapter presentations, the perceptive shift observed after 95 adapter presentations was not significantly bigger than the shift observed after 32 presentations.

Recalibration and other perceptual aftereffects tend to dissipate over time (Epstein, 1975; Hershenson, 1989). However, lexically-driven perceptual learning effect observed by Eisner and McQueen,(2006) remained stable over a long period of time. The aftereffects were measured directly after exposure to an ambiguous phoneme which was embedded in a lexical context, and again 12 hours later. Remarkably, the magnitude of the observed aftereffect was not smaller 12 hours after exposure, showing no sign of dissipation whatsoever. Although the time period was much shorter, (i.e. 25 minutes) Kraljic and Samuel (2005) reported similar results as they reported lexically-driven perceptual learning to remain stable even after various unlearning procedures. In chapters 3 and 4 we investigate the rate of dissipation of lipread-driven recalibration, and compare it with the dissipation of selective speech adaptation. Chapter 6 provides an explorative study on lipread versus lexical recalibration and their stability over time.

Developmental trend

It is generally claimed that perceptual learning of speech becomes more difficult with age. Phoneme-boundaries become less flexible with age and adults should have more difficulty acquiring new phonemic contrasts than children. A decrease of the neural plasticity with age is supposedly the reason for this decrease in flexibility (Gilbert, 1994; Gilbert et al., 2001; Lenneberg, 1967). Recently however, an alternative theory was proposed, stating that future learning is inhibited by prior learning. Already formed phoneme categories interfere with acquiring new categories and therefore complicate learning and adaptation. In this respect, it is not so much the age of the listener that is relevant for perceptual learning in speech, but the language experience itself (Kuhl, 2000). Nevertheless, both theories would predict a negative relation between the magnitude of recalibration in speech perception and age.

As for audiovisual speech perception, despite demonstrations that lipreading contributes to speech perception in early infancy, it has also been found that the impact of lipreading increases with age. In the original study by McGurk and MacDonald, (1976) a group of adult participants (18 – 40 yr), and two groups of children aged 3-4 and 7-8 years were presented audio-visual incongruent speech tokens. Results showed that adults were more influenced by the incongruent visual information than the two younger age groups, while the two younger groups did not differ consistently. Such a developmental trend has been confirmed in later studies (Sekiyama, 2003; Burnham, 2004). The literature on perceptual learning and audiovisual speech perception therefore predict opposite results for a developmental trend in visual recalibration with increasing age. In chapter 5, we explored this issue by investigating visual recalibration in two age groups, 5-year olds and 8-year olds, providing the first data regarding a developmental trend.

Neural processing of audiovisual speech

Integration of information provided by two or more modalities takes place in so-called multimodal brain areas (Stein & Meredith, 1993). A multisensory brain region contains neurons which are responsive to the input of more than one modality. In a detailed study recording the activation of single cells in the superior colliculus (SC) of the cat, Stein and Meredith (1993) described neurons whose activity was significantly increased with the stimulation of audiovisual inputs which were temporally and spatially aligned. With bimodal stimulation, the firing rate of these cells was substantially higher than the summation of the unimodal responses, i.e. these cells demonstrated a supra-additive response to multisensory stimulation. The supra-additive response is considered to be a new, genuine multimodal response, signifying the integration of bimodal stimulation. The enhancement of activation to multimodal stimulation is maximal when the neural

response to a single sensory input is minimal, a principle now known as inverse-infectiveness. When however the auditory and visual information were spatiotemporal incongruent, the response of multisensory neurons was decreased as opposed to unimodal stimulation. Subsequent brain imaging and electrophysiological studies adopted the observed supra-additive and inverse-infectiveness in the CS to as described by Stein and Meredith to define areas of multisensory integration in the brain.

In monkey studies and human brain imaging studies, sites of multisensory integration have now been identified throughout the brain, in both cortical and sub cortical areas (see Calvert et al., 2004 for a review), demonstrating that combining and utilizing information from more than one modality is a general function of the nervous system.

Integration of lipread information and auditory speech

The brain area which has been most often reported as the site of audiovisual speech integration is the left superior temporal sulcus (STS) and the left superior temporal gyrus (STG) (Callan et. al, 2003; Calvert et al. 2000; Macaluso et al., 2004). Several functional magnetic resonance imaging (fMRI) and electrophysiological studies demonstrated a modulation in activity of the STS during auditory, visual and audiovisual speech perception, suggesting a role for the STS in the integration of acoustic and visual speech cues (Beachamp et al., 2004; Calvert et al., 2000; Campbell et al., 2001; Sekiyama et al., 2003b). In an fMRI study by Calvert et al. (2000) participants were presented with visual, auditory and audiovisual speech cues. The bimodal conditions could either be congruent or incongruent (e.g. the lip gestures were either matching or non-matching to the auditory speech). The only brain area to meet the three criteria of multi-sensory integration was the ventral bank of the STS in the left hemisphere. Firstly, this area was responsive to both auditory-only and visual-only presented speech, as a multimodal area should be responsive to input from more than one modality (Stein & Meredith, 1993). Furthermore in the left STS, supra- and sub-additive responses were observed, produced by exposure to congruent and incongruent stimuli respectively. When presented with congruent audiovisual speech a supra-additive enhancement of response of 30 – 80% above the summation of the unimodal responses was observed in this area. Incongruent audiovisual speech resulted in a sub-additive response in the left STS by 50% of the summed unimodal responses: These results suggest that the STS plays an important role in the integration of auditory and visual speech cues. A response-enhancement for bimodal as opposed to unimodal speech was also observed in the primary auditory and visual motor cortices. The authors suggested that the auditory and visual speech cues are integrated in the left STS and then modulate processing in the sensory-specific cortices through subsequent back-projections.

Sams et. al (1991) were the first to report that neuromagnetic auditory mismatch response (MMNm) can be triggered by an illusory auditory change driven by lipread information. The MMNm is the magnetic counterpart of the electrophysiological mismatch negativity (MMN) which is an auditory brain potential, evoked by a discriminable change in an otherwise consistent auditory stimulus train. Sams et al. presented participants with two audiovisual speech tokens which only differed in the visual component but were perceived differently due to the McGurk-effect. The incongruent stimulus causing a McGurk effect was auditory /pa/ combined with a face articulating /ka/ inducing the percept /ta/ or /ka/. The second stimulus was audiovisual congruent /pa/. Both audiovisual stimuli served as the standard or the deviant stimulus. As a MMN-response was elicited by the deviant stimuli which only differed from the standard by visual information, it demonstrated that acoustically similar but perceptually different stimuli are processed differently in the auditory cortex, (see Colin et al., 2004; see Möttönen et al., 2002; Saint-Amour et al., 2007 for similar results)

Other electrophysiological studies reported very early effects of visual speech on the processing of auditory speech in the primary auditory cortex (Möttönen et al., 2002; V. van Wassenhove et al., 2003; V. van Wassenhove et al., 2005). Van Wassenhoven et al. (2003) reported that the presentation of congruent and incongruent lip gestures while listening to nonsense syllables reduced the amplitude of the auditory evoked N1/P2 complex and temporally facilitated the P1/N1/P2 complex as compared to the presentation of the auditory-only stimuli. Visual speech thus facilitated auditory speech perception in the auditory cortex already after 50 – 100 ms post-auditory onset.

Together, behavioural, electrophysiological and fMRI studies all provide evidence that lipread information affects speech processing at an early perceptual stage, possibly before phonetic processing. In addition to the STS, greater responses to audiovisual congruent speech than for unimodal auditory or unimodal visual speech, has also been observed in the primary auditory and visual motion cortices (Calvert et al., 1999; 2000) As a plausible mechanism of audiovisual integration of speech, Calvert et al. (1999; 2000) suggested that the multisensory speech input is initially integrated in the STS/ STG, and that this area subsequently activates and modulates processing in the sensory specific auditory association and visual cortices through feedback projections.

Lexical feedback in models of speech perception

The discussion regarding the nature of lexical effects on speech perception emanates from the still unresolved and current issue whether speech perception occurs in either an autonomous or interactive process (McClelland et al., 2006; McQueen et al., 2006). Most models on spoken word perception incorporate at least two levels of

processing. In the first prelexical stage, the acoustic signal is converted into an abstract description of a speech utterance (i.e. a phonetic representation) and in the lexical stage; this abstract description is combined with stored lexical and semantic knowledge in order to add meaning to the sounds. A long standing issue regards the question whether speech perception is an interactive process, allowing feedback from the lexical level onto processing at the prelexical level or whether the processing of speech sounds is an autonomous, feed-forward system (McQueen et al., 2006; Norris et al., 2000).

Proponents of interactive models of speech perception like TRACE (McClelland & Elman, 1986) argue that the online effects of lexical information as occurring in the Ganong-effect are perceptual in nature and arise from feedback projections from the lexical onto the prelexical level. Interactive accounts thus predict a flow of information from the lexical level directly onto a prelexical level of processing which than actually changes the perception of speech sounds.

The TRACE model (McClelland & Elman, 1986) consists of a network of three interconnected layers of processing: an acoustic or phonetic feature layer, a phoneme layer and a word layer. Each layer consists of a number of units, representing the particular speech elements relevant for that layer. In the phoneme layer for example, there is a separate unit for each phoneme which is possibly present in the speech input; the feature layer consists of units representing for example voicing and burst. The three layers have bidirectional excitatory connections and within each layer there are competitive inhibitory connections between the units. Through the bidirectional excitatory connections, the model thus incorporates a feedback mechanism for online speech perception from higher lexical levels (the word layer) onto lower levels of speech processing. Online lexical effects on phoneme perception are thus explained by these feedback projections. The model might account for effects of perceptual learning as the network updates itself by strengthening or retuning the connections between units which were simultaneously activated. Lexical information can reach down and retune prelexical mechanisms, thereby mediating boundary adjustment (McClelland et al., 2006).

Autonomous approaches state that speech processing is a strictly feed-forward process, in which there is no feedback from lexical levels onto lower levels of processing. In fact, as stated by Norris et al. (2000) feedback from lexical onto prelexical processing stages is not beneficial, as lexical feedback would only confirm phonemic decisions which are already made at the phoneme level. Moreover feedback connections potentially impair correct perception as lexical information can overwrite information present in the input. Norris et al. presented the Merge-model that consists of three levels of processing: input nodes, lexical nodes and decision nodes. Information proceeds from the prelexical level onto the lexical level, but there is no feedback from the lexical onto the prelexical level. The decision nodes however, are sensitive to both lexical and prelexical information and it

is there that an explicit decision is made regarding the identity of the speech input (Norris et al., 2000). Integration of information from the prelexical input nodes and the lexical nodes thus takes place at the post-perceptual decision-stage. The effects of lexical information on online speech perception arise at the post-perceptual decision level. Although this model allows no online feedback from the lexicon onto prelexical stages, it was suggested by Norris et al. (2003) that off-line feedback for learning still is possible in Merge. Recalibration of phonemic representations is then acquired by means of off-line feedback for learning from lexical information onto prelexical phoneme representations.

The study by Samuel (1997) reporting selective speech after exposure to lexically-restored phonemes, was taken to be a clear demonstration of lexical feedback onto lower prelexical levels of processing, an observation which would only be attributable to feedback from the lexical onto the prelexical level. It could however also be argued that as the experiment by Samuel (1997) does not measure an online effect of lexical information on phoneme processing, that other effects came into play. Possibly, over time, the perceptual system has learned to incorporate the noise sound as being an exemplar of either the /b/ or /d/ phoneme category. Prolonged exposure to this sound would then produce selective speech adaptation as a result of this off-line learning; not as a result of online feedback effect from the lexical level onto a prelexical processing stage.

Autonomous and interactive modes might thus both incorporate lexically-guided recalibration of phoneme identification, but they suggest different mechanisms. The main difference between TRACE and Merge is, that according to TRACE, lexical effects are perceptual in nature and arise from direct lexical feedback onto the phoneme layer. In Merge, direct bias effects arise at a post-perceptual decision stage, as lexical and prelexical information affects phoneme identification only at the decision nodes.

The work presented in the present thesis aims to expand the knowledge on the processes of lexical- and lipread-driven recalibration; thereby improving our understanding of the mechanisms and processes on which these learning effects are based.

Overview of the thesis

The first part of the thesis, (chapters 2, 3, 4, 5 and 6), is mainly concerned with selective speech adaptation and lipread driven recalibration. Effects of lipread recalibration are investigated by exposing participants to audiovisual speech tokens, containing an ambiguous speech sound between /aba/ and /ada/ which is dubbed onto a face articulating either /aba/ or /ada/ (hence A?Vb or A?Vd). The occurrence of selective speech adaptation is investigated by exposing participants to an unambiguous speech sound dubbed onto a face making congruent articulatory gestures (AbVb and AdVd). Aftereffects produced by the exposure are measured on subsequent auditory-only identification trials, consisting of ambiguous speech tokens of the /aba/ - /ada/ continuum. We investigated the roles of auditory and visual speech information for the occurrence of selective speech adaptation and recalibration. Secondly, in order to demonstrate the existence of dissociation between these two aftereffects, their dissipation rates and build-up courses are investigated. Chapter 6 investigates the developmental trend of both aftereffects.

In the second part of the thesis, lipread and lexical driven recalibration on phoneme identification are investigated and compared. New stimuli were created which enabling us to test lipread and lexical driven recalibration on the same ambiguous phoneme. An ambiguous phoneme between /p/ and /t/ were spliced after words normally ending in /p/ or /t/, as for example vloot (meaning *fleet* in Dutch) and hoop (meaning *hope* in Dutch) in order to investigate lexically driven recalibration; the same ambiguous phoneme was also spliced after pseudowords and dubbed onto a face articulating that pseudoword ending in a /p/ or /t/ (for example: woop and woot) in order to investigate lipread driven recalibration. In chapter 7, the effects of disambiguating lipread and lexical information on phoneme perception are investigated and compared. Chapter 8 provides an electrophysiological study of lexically driven recalibration, which investigates the extent to which this effect is perceptual in nature.

The question addressed in chapter 2 is whether recalibration can be induced with a McGurk stimulus, containing an auditory clear phoneme instead of an auditory ambiguous one. Two strategies are adopted to disentangle these two aftereffects from each other. One strategy is to vary the number of exposure stimuli. It is known that selective speech adaptation increases with increasing number of exposure (Ohde & Sharf, 1979) but for recalibration no data regarding build-up were present yet. As a second strategy we introduced auditory clear phonemes during posttests, as it might be possible that recalibration induced by a McGurk stimulus is token-specific (Kraljic & Samuel, 2006; Eisner & McQueen, 2005) would than only be observable on the auditory clear phoneme. The main conclusion drawn from these experiments is that for the occurrence of recalibration, the combination of an ambiguous speech sound with lipread speech is

essential. Exposure to stimuli containing a clear auditory speech token always produces selective adaptation to speech with no role for visual speech information.

In chapter 3 and 4 the rate of dissipation of selective speech adaptation and recalibration are compared. The respective rates of dissipation are investigated as a function of posttesting period and in chapter 4 the effect of exposure duration on the stability of selective adaptation and recalibration is investigated. Recalibration and selective speech adaptation were clearly dissociated as the two effects dissipate at different rates. Whereas selective speech adaptation shows no sign of dissipation, the effect of recalibration dissipates fast with prolonged post-testing. Increasing exposure duration did not affect the dissipation rates of selective speech adaptation or recalibration.

The build-up courses of recalibration and selective speech adaptation are investigated and compared in chapter 5. The number of adapter presentations is cumulatively increased, and auditory-only posttests are interspersed after 1, 2, 4, 8, 16, 32, 64, 128, and 256 adapter presentations in order to measure the build-up course. The results for the audiovisual ambiguous adapters are somehow surprising since the recalibration effect initially increases, then remains stable, but then decreases when more adapters were presented. In contrast, selective speech adaptation induced by exposure to audiovisual congruent adapters increases linearly with increasing number of adapter presentations.

The influence of visual information on auditory speech perception increases with age, Sekiyama, 2003; Burnham, 2004; McGurk & MacDonald, 1976). The study presented in chapter 5 further explores the developmental trends in the use of lipread information. So far, all previous studies looked at the immediate effect that a lipread stimulus has on a simultaneously presented auditory speech token. Here, we tested whether there is, as in the case of immediate visual bias, a developmental trend in the use of visual speech for the purpose of recalibration. Results indicate that there is indeed an increase of the recalibration effect with increasing age in children.

In chapter 7 the question is addressed whether the phonetic adjustment - or recalibration – differs when it is evoked by lipread information versus lexical knowledge. Lipread and lexically driven recalibration are investigated in terms of magnitude, rate of dissipation, and stability over time. We also investigate for a correlation between direct bias and recalibration. The results provided strong indications that there is no fundamental difference between the use of top-down lipread information and bottom-up lexically information for the recalibration of phoneme boundaries. Whether the disambiguating information is lipread speech or lexical stored knowledge, the magnitude of the aftereffect, or its rate of dissipation are similar.

Lexical information can bias categorization of an ambiguous phoneme and subsequently evoke a shift in the phonetic boundary; the electrophysiological study

presented in chapter 8 explores the extent to which the lexical driven recalibration is perceptual in nature. We investigated whether the lexically driven adjustment in the perception of an ambiguous phoneme, is reflected on the auditory evoked mismatch negativity (MMN). To this end, auditory stimuli are presented in a typical odd-ball sequence in which the standard was an ambiguous sound halfway between /t/ and /p/ presented at the end of a Dutch word normally ending in /t/ ('vloot', meaning "fleet") or /p/ ('hoop', meaning "hope"). The non-ambiguous sound /t/ embedded in the same context served as the deviant stimulus. The amplitude of the MMN-response, indexing the perceptual difference between the ambiguous sound and unambiguous /t/ was bigger for the p-word 'hoop' than the t-word 'vloot'. This result was taken to indicate that lexical information actually reached down to early perceptual processing stages.

Chapter 2

Visual recalibration versus selective adaptation of auditory speech: The role of sound ambiguity.

Introduction

A major part of research on multisensory processing has been carried out with conflict situations, in which incongruent information about potentially the same distal event is presented to different modalities (e.g., de Gelder & Bertelson, 2003). Exposure to such conflicting inputs produces two main effects: immediate biases and aftereffects. By immediate biases are meant effects of incongruent inputs in a distracting modality on the perception of corresponding inputs in a target modality. For example, in the so-called ventriloquist illusion, the perceived location of target sounds is displaced toward light flashes delivered simultaneously at some distance, in spite of instructions to ignore the latter (Bertelson, 1999). Aftereffects are shifts observed following exposure to an intersensory conflict, when data in one or in both modalities are later presented alone. For the ventriloquism situation, unimodal sound localization responses are, after a period of synchronized exposure to spatially discordant sound bursts and light flashes, shifted in the direction of the distracting flashes. The occurrence of aftereffects has generally been taken as implying that exposure to incongruence between corresponding inputs in different modalities recalibrates processing in one or both modalities in a way that eliminates (or at least reduces) the perceived discordance (Welch, 1978b).

Although immediate biases and aftereffects have consistently been demonstrated for spatial conflicts, the existing evidence has long been less complete for conflicts regarding event identities. Here, biases were often reported, but, for some time, no aftereffects indicative of recalibration. The main example is the conflict resulting from the acoustic delivery of a particular speech token in synchrony with the optical presentation of a face articulating a visually incongruent token. As originally reported by McGurk and MacDonald (1976), this kind of situation generally produces strong immediate biases of the auditory percept towards the speechread distracter. This phenomenon is now generally called 'the McGurk effect'. On the other hand, no demonstration of recalibration consequent upon exposure to McGurk situations was for some time reported. Roberts and Summerfield (1981) exposed participants to incongruent pairs consisting of auditory tokens /bɛ/ and /dɛ/ synchronized with the visual presentation of a face articulating /gɛ/ (pairs AbVg), as well as to unimodal auditory speech tokens /bɛ/ and /dɛ/. Exposure to these unimodal adapters resulted in a reduction in the proportion, during subsequent identification tests, of responses consistent with the exposure auditory tokens. This is the phenomenon known as selective speech adaptation to speech (Eimas & Corbit, 1973; Samuel, 1986b) Regarding recalibration, the important finding was that exposure to the bimodal pair AbVg produced aftereffects that were no bigger than those produced by the unimodal Ab. The same lack of effect of an incongruent visual token was later replicated by Saldaña and Rosenblum (1994) and taken as possibly revealing a basic difference between identity and spatial conflicts (Rosenblum, 1994).

Starting from the hypothesis that the preceding failures to obtain visual recalibration could somehow have been caused by the selective speech adaptation consequent on repeated exposure to end-point auditory tokens, Bertelson et al. (2003) examined what would happen if these end-point tokens were replaced by a single ambiguous token combined with the same visual stimulus. They conducted an experiment in which participants were exposed to bimodal pairs in which the auditory component was each participant's most ambiguous speech token from an /aba/ - /ada/ continuum (henceforth A?), and the visual component featured the unambiguous articulation of either of the two end points, /aba/ or /ada/. Following the habitual conflict adaptation paradigm, auditory identification tests, using as material the ambiguous token A? (and two slightly less ambiguous ones) were administered after exposure to bimodal adapters A?Vb or A?Vd. In this situation, recalibration occurred: /aba/ responses were more frequent after exposure to A?Vb than after exposure to A?Vd.

As a check that the success in demonstrating visual recalibration was effectively linked to the use of an auditory ambiguous token during exposure, in a second experiment the same two visual adapters (Vb and Vd) were combined with congruent non-ambiguous auditory tokens, giving adapting pairs AbVb and AdVd. Exposure to these pairs resulted in reductions of the proportion of responses consistent with the bimodal adapter, i. e. selective adaptation. Fewer /aba/ responses were given after exposure to AbVb than after exposure to AdVd, an outcome opposite the one obtained with the ambiguous auditory adapter A?. Interestingly, in subsequent identification tests, corresponding exposure pairs with ambiguous versus non-ambiguous auditory component (AbVb vs. A?Vb and AdVd vs. A?Vd) were judged identically, ruling out any contribution of deliberate strategies or response biases to the contrasting adaptation effects.

In the present experiment, the relations between visual recalibration and selective speech adaptation are further explored. One question of concern was the reason why in the studies of Roberts and Summerfield (1981) and of Saldaña and Rosenblum (1994) exposure to McGurk pairs with non-ambiguous auditory tokens combined with incongruent visual tokens (i.e. AbVd or AdVb) did not produce larger aftereffects than exposure to unimodal auditory tokens (Ab or Ad). As was noted by Bertelson et al. (2003) exposure to McGurk pairs is expected to produce both selective speech adaptation and visual recalibration, while exposure to auditory-only adapters would produce only selective adaptation. One possible reason why the additional effect of recalibration was not found in the earlier studies was provided in the study by Bertelson et al. (2003) was, "that same-direction effects produced by separate adaptation processes do not sum". To check on the reality of the no-summation result, both McGurk exposure conditions AbVd and AdVb, and the corresponding unimodal auditory exposure conditions Ab and Ad were included in the present study. Also included were two unimodal visual exposure conditions, Vb and Vd, to

insure against the possibility that visual tokens would by themselves produce selective adaptation. Such visual exposure conditions were run by Roberts and Summerfield (1981) and exerted no effects on subsequent auditory identification tests.

Another innovation of the present study is that effects of number of exposure trials were considered. Whereas the original recalibration study was run with 8-exposures adaptation blocks, both 8-exposure and 50-exposure blocks were run, in all the conditions of the present experiment. Regarding the original conditions used in the study by Bertelson, the consideration of different exposure lengths provided a useful check on the robustness of the contrasting effects. It provided an opportunity to explore the possibility that the ambiguity of the auditory adapter would affect, beyond the direction of the adaptation effects, also the development of their respective build-ups. The consideration of different adaptation block sizes proved important on the other hand for the interpretation of the aftereffects obtained with the McGurk pairs. This point will be developed in the general discussion.

Experiment 1

Method

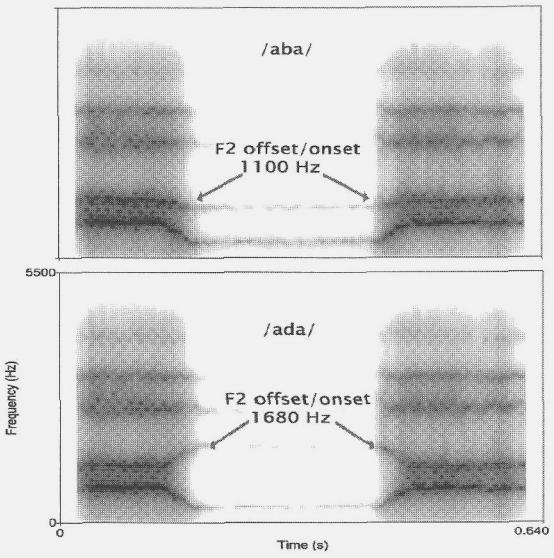
Participants

Fourteen first-year students from Tilburg University (4 male, 10 female) participated in three sessions each. All were naïve as to the purpose of the experiment.

Materials

Details of the stimuli have been described elsewhere (Bertelson et al., 2003; Vroomen et al., 2004). In short, a 9-point /aba/-/ada/ speech continuum was created by varying the frequency of the second (F2) formant in equal steps. The endpoints and the most ambiguous auditory tokens (see figure 1) were dubbed onto the video of a face articulating /aba/ or /ada/.

Figure 1



Spectrogram of the /aba/ - /ada/end-points.

Procedure

The experimental testing comprised successively a calibration phase, a training phase an adaptation phase consisting of a series of adaptation blocks, each with exposure trials followed by post-tests. Testing ended with a stimulus identification phase.

The calibration phase served to determine each participant's individual /aba/-/ada/ 50% cross-over point. The participant gave forced-choice /aba/ or /ada/ key pressing responses to presentations of each of the nine items of the auditory continuum. Ninety-eight trials were presented in random order at 1.5 s intervals. Tokens from the middle of the continuum were presented more often than tokens from extreme locations (6, 8, 14, 14, 14, 14, 14, 8 and 6 presentations for each of the nine tokens, respectively). The continuum token closest to the participant's phoneme boundary was chosen as the participant's most ambiguous auditory token (henceforth A?) for the subsequent testing.

On the 60 trials of the training phase, the same forced-choice responses were given to 20 presentations, in random order, of the ambiguous auditory token A?, and of its two nearest continuum neighbors, A?-1 and A?+1.

Adaptation blocks each consisted of either 8 or 50 exposure trials, presented at 600 ms intervals, followed by six auditory post-tests, namely each of the three tokens A?-

1, A? and A?+1 presented twice in randomized order. Exposure trials were run under five different conditions: 1).The audiovisual ambiguous condition, with each trial involving A? combined with either visual /aba/ (pair A?Vb) or with visual /ada/ (pair A?Vd); 2) The audiovisual non-ambiguous condition, with pairs AbVb and AdVd; 3) The audiovisual incongruent condition, with pairs AbVd and AdVb, 4). The auditory- only condition, with presentations of auditory /aba/ (Ab) or /ada/ (Ad); 5) the visual-only condition, with presentations of either visual /aba/ (Vb) or /ada/ (Vd). No overt response was required during exposure, but to ensure attention to the face, the participant was instructed (in conditions 1, 2, 3 and 5) to report (by key press) presentations of a small (12 pixels) white spot, 100 ms in duration, between the nose and the upper lip. During exposure a total of 200 of such catch trials were delivered at unpredictable times.

For each length of exposure (8 and 50), 10 different adaptation blocks (5 exposure conditions x 2 tokens (/b/ or /d/) were each run five times in counterbalanced order. Half the participants were run first on the 8-exposure blocks, then on the 50 exposure blocks, and the other ones had the reversed order. Total testing lasted about 3 hours, distributed over the three sessions. At the end of the experiment, forced-choice identification tests was run to check whether the auditory endpoint tokens were indeed perceptually influenced by the incongruent articulatory gestures. The audiovisual congruent and audiovisual incongruent stimuli were presented 20 times in randomized order (2.5 s. ISI). Participants were instructed to judge the stimuli either as /aba/, /ada/ or /abda/.

Results and discussion

The average percentage of /b/ responses during training was 46%, indicating that the two response alternatives (/b/ or /d/) were about equally distributed when presented in the context of posttest stimuli. Identification scores of the exposure stimuli showed that perception of the audiovisual incongruent stimuli was indeed visually biased, see table 1. 94% of the catch trials were detected, showing that participants indeed attended the videos during the exposure phase.

The adaptation relevant results appear in table 2. For each exposure condition, it shows the percentage /aba/ responses on post-tests for each exposure alternative, and the resulting mean aftereffects. The latter were calculated by subtracting the percentage of /aba/-responses with /ada/ adapters from those with /aba/ adapters. Aftereffects were calculated (as in Bertelson et al., 2003), by the difference in the proportion of /aba/ responses obtained after exposure to respectively A?Vb and A?Vd (ambiguous sound condition) or after AbVb and AdVd (non-ambiguous sound condition), or Ab and Ad (auditory only condition or Vb and Vd (visual only condition). Negative aftereffects mean

fewer responses consistent with the adapter and reflect selective adaptation; positive aftereffects mean more consistent responses, or recalibration.

The main fact apparent in the table is that negative aftereffects occurred after exposure to non-ambiguous auditory tokens, whether combined with congruent, incongruent or no visual tokens, while positive aftereffects were observed only in the audiovisual ambiguous condition. In consequence, we submitted the data (aftereffects) for the three conditions with non-ambiguous auditory tokens to a MANOVA with Visual Component (Congruent, Incongruent, None) and Exposures (8, 50) as factors. Exposure produced the only significant effect, $F(1,13) = 19.70$, $p < .001$. The main effect of Visual Component, $F(2,26) < 1$, and its interaction with Exposure, $F(2,26) = 1.53$, $p = .234$, were non-significant.

Separate t-tests showed that all adapters produced aftereffects significantly different from zero (all p-values at least < .015), except the visual-only ones (both p's > .10). Aftereffects were enhanced with 50 exposures compared to 8 exposures for the audiovisual congruent condition, $t(1,13) = 2.77$, $p < .02$, the audiovisual incongruent condition, $t(1,13) = 2.56$, $p < .025$, and the auditory-only condition, $t(1,13) = 4.98$, $p < .001$, but there was no effect for the number of exposure repetitions in the audiovisual auditory-ambiguous condition, $t(1,13) = 1.02$, $p = .33$, or in the visual-only condition, $t(1,13) = .742$, $p = .472$.

Table 1.
Stimulus Identification in Experiment 1.

		Response		
Stimulus Condition		/aba/	/ada/	/abda/
Audiovisual congruent	AbVb	.99	.003	.007
	AdVd	.02	.96	.025
Audiovisual incongruent	AbVd	.75	.19	.053
	AdVb	.03	.41	.57

Values represent response proportions.

Table 2.
Mean Percentage of /aba/ Responses for the Exposure Conditions in Experiment 1.

Exposure Condition	/b/-Adapter		/d/-Adapter		Aftereffect
8 Exposures					
Audiovisual congruent	AbVb	32	AdVd	49	-17*
Auditory ambiguous	A?Vb	41	A?Vd	29	+14*
Audiovisual incongruent	AbVd	32	AdVb	52	-20*
Auditory-only	Ab	31	Ad	46	-15*
Visual-only	Vb	40	Vd	37	+3
50 Exposures					
Audiovisual congruent	AbVb	24	AdVd	56	-32*
Auditory ambiguous	A?Vb	46	A?Vd	30	+16*
Audiovisual incongruent	AbVd	26	AdVb	64	-38*
Auditory-only	Ab	19	Ad	61	-42*
Visual-only	Vb	47	Vd	41	+6

Note: Aftereffects were calculated by subtracting the percentage of /aba/-responses following exposure to /d/-adapters from corresponding /b/-adapters. Negative aftereffects denote fewer responses consistent with the adapter, whereas positive aftereffects reflect more responses consistent with the adapter.

The present results dissociate recalibration and selective speech adaptation in two ways: in the direction of the aftereffects and secondly in the effect an increase in the number of adapter presentations has on the magnitude of the effect. Exposure to adapters containing an auditory clear speech token decreased the tendency to report that token on subsequent posttests, regardless of the presence or identity of visual information. All these adapters thus produced selective speech adaptation to speech and did so equally strong. In addition, with increasing the number of presentations to these adapters, the magnitude of the selective adaptation-effect increased, also equally strong for all adapters. Thus, the audiovisual incongruent adapters affected auditory speech perception on post-tests similarly as the audiovisual congruent or auditory-only adapters.

In contrast, exposure to the audiovisual ambiguous adapters increased the proportion visual consistent responses on posttests showing recalibration. Contrary to selective adaptation, the magnitude of recalibration did not increase when more adapters were presented.

Experiment 2

Although it is possible that exposure to audiovisual incongruent stimuli does not produce recalibration, there are two alternative explanations for the results obtained in Experiment 1.

In the study by Bertelson et al., (2003) the biggest recalibration effects were found on the most ambiguous testtoken. This may have occurred for two reasons: either the boundary between two phonemes is shifted, or alternatively, because the specific token present during exposure is assigned to a new category. In the latter case, one expects recalibration to be token-specific (see Eisner & McQueen, 2005). Importantly, in the study by Bertelson et al. (2003) the ambiguous token was the auditory component of the audiovisual adapter presented during the exposure phase. It might be that the auditory component of the audiovisual stimulus presented during the exposure phase is the auditory token which is recalibrated, in which case recalibration effects will be bigger on the auditory token which was presented during the exposure phase. This is indeed the procedure used for the audiovisual ambiguous exposure stimuli, but not for the audiovisual incongruent stimuli, as aftereffects by exposure to audiovisual stimuli containing a clear auditory token were measured on auditory ambiguous identification trials. This procedure assumes possible recalibration effects to generalize to phonemes which were not presented during the exposure phase.

A second problem relates to the perception of the audiovisual incongruent stimuli. Since the audiovisual stimulus AdVb is often perceived as /abda/ (see table 1), recalibration by exposure to AdVb might therefore reveal itself by an increase the proportion /abda/, and not /aba/ responses, on posttest trials.

To control for these two alternative explanations, during the second experiment, posttests contained stimuli from whole range of the /aba/ - /ada/ continuum and we added /abda/ as a third response option during post-testing.

Method

Participants

20 new participants engaged in Experiment 2. All were first-year psychology students and received course-credit for their participation.

Materials

The materials were the same as used in Experiment 1.

Procedure.

Testing consisted of a baseline test and an exposure-posttest phase. During baseline testing, participants were presented with the tokens 1, 3, 5, 7, and 9 of the 9-point /aba/ - /ada/ continuum. All tokens were presented 20 times in randomized order, (2 s. ITI) and had to be identified either as /aba/, /ada/ or /abda/ by pressing a corresponding key on the keyboard.

The adaptation blocks consisted of 8 exposure trials, presented at 600 ms intervals, immediately followed by 5 auditory-only posttests, which were to be identified as /aba/, /ada/ or /abda/. Exposure trials were run under three conditions: audiovisual congruent (AbVb, AdVd), audiovisual incongruent (AbVd, AdVb) and audiovisual ambiguous,(A?Vb A?Vd).The six different exposure block (3 exposure conditions x 2 tokens, (/b/ or /d/)) were each presented ten times. Presentation order of the exposure blocks was counterbalanced. Posttests consisted of the presentation of auditory tokens 1, 3, 5, 7, and 9 of the 9-point /aba/ - /ada/ continuum, which had to be identified as /aba/, /ada/ or /abda/. Presentation order of auditory-only testtrials was counterbalanced over conditions. During exposure, participants were instructed to silently watch the videos and press a special key upon detecting a catch trial. Total testing lasted about 2 hours, including a short pause halfway the experiment.

Results and discussion

Participants detected 95% of the catch trials on average.

Within each condition, aftereffects were calculated for each testtoken (1, 3, 5, 7, and 9) and response category (/aba/, /ada/ and /abda/) separately, such that a negative sign reflects selective speech adaptation and a positive sign reflects recalibration. For example, the percentage /aba/ response on testtoken 1 after AbVb exposure was subtracted from the percentage /aba/ response on testtoken 1 after AdVd exposure, signifying the difference in /aba/ response after AbVb exposure as opposed to AdVd exposure on testtoken 1. Aftereffects are displayed in the figures 2 to 4. Inspection of these figures shows that the aftereffects obtained for each auditory testtoken are small and highly similar for all exposure conditions.

An overall 3 (Exposure Condition: audiovisual congruent, audiovisual incongruent, and audiovisual ambiguous) x 5 (Auditory Testtoken 1, 3, 5, 7 and 9) x 3 (/aba/, /ada/ or /abda/ Response) ANOVA was run over the aftereffects. Only the 3-way interaction Exposure Condition by Auditory Testtoken by Response was significant, $F(16, 304) = 1.88$, $p < .025$. None of the other main or interaction effects was significant, (all p-values > .15).

For each of three audiovisual conditions a 5 (Auditory Testtoken) x 3 (Response) ANOVA was run. Only for the audiovisual incongruent exposure stimuli was the interaction

between auditory testtoken and response was significant $F(8, 152) = 4.43$, $p < .001$. None of the other main- or interaction effects was significant, (all p-values > .1). One-sample t-tests showed increase in /aba/ responses after AdVb as compared to AbVd on testtoken 5, was significant, $t(1,19) = 4.22$, $p < .001$.

No evidence for recalibration induced by an audiovisual incongruent speech stimulus was found. Although visual information does influence the online perception of a clear auditory speech token, exposure to a McGurk stimulus does not seem to cause recalibration on the perception of that auditory clear speech token.

Contrary to the results observed in Experiment 1, none of the conditions produced significant aftereffects. Response patterns on posttests were similar for all conditions. Probably, this result is caused by the option to respond /abda/ during the posttest trials. After finishing the experiment, most participants reported that sometimes they found it hard to identify testtokens either as /aba/ or /ada/, in which case they responded /abda/ in spite of the explicit instruction only to respond /abda/ when two successive phonemes /b/ and /d/ were perceived. Consequently, the proportion /abda/ responses was highest on auditory ambiguous testtokens, implying that when participants were not completely sure regarding the identity of the phoneme being either /aba/ or /ada/, they avoided making this choice by responding /abda/. In spite the fact that the midpoint tokens are never as clearly perceived as the continuum-endpoint tokens, Although perception of the ambiguous phoneme could still be biased towards hearing either more /aba/ or /ada/. In the present experiment, the contrast between ambiguous and clear speech token is even more obvious, as the clear endpoint tokens were also present in the posttest. The option to respond /abda/ in the present experiment thus probably contaminated the data.

To summarize, although the most important result here is that the alternative explanations as to the question why no recalibration was observed following audiovisual incongruent can be excluded. It is therefore concluded that the McGurk stimuli did not produce recalibration on the endpoint auditory phonemes. Furthermore, the failure to observe recalibration by the McGurk-stimuli is not due to the presentation of only auditory ambiguous tokens in posttests.

Figure 2

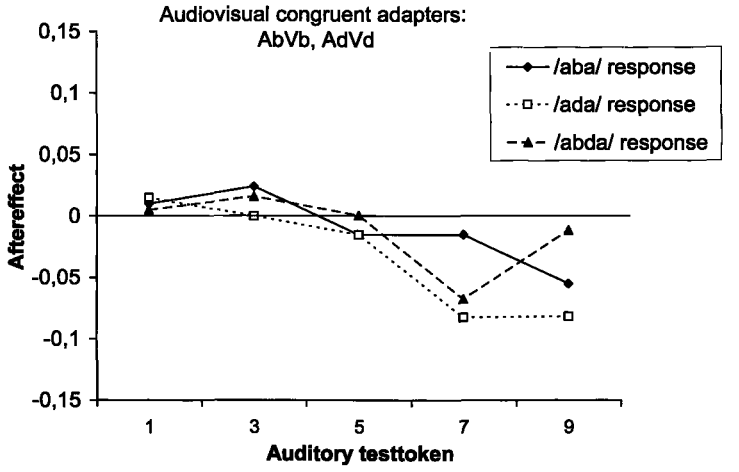


Figure 3

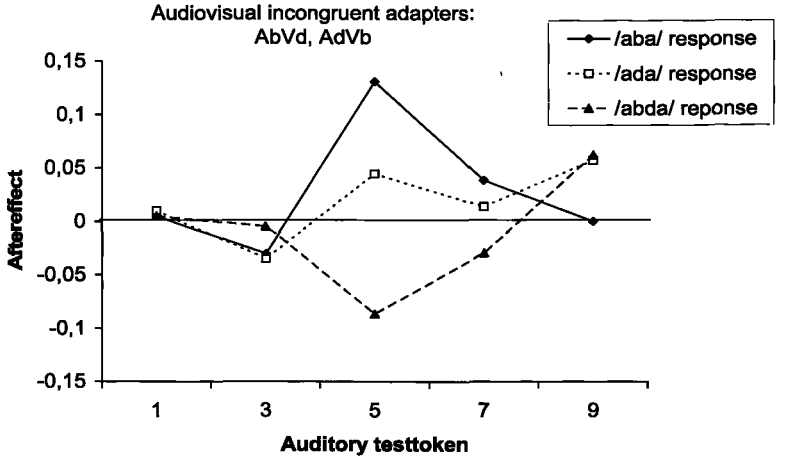
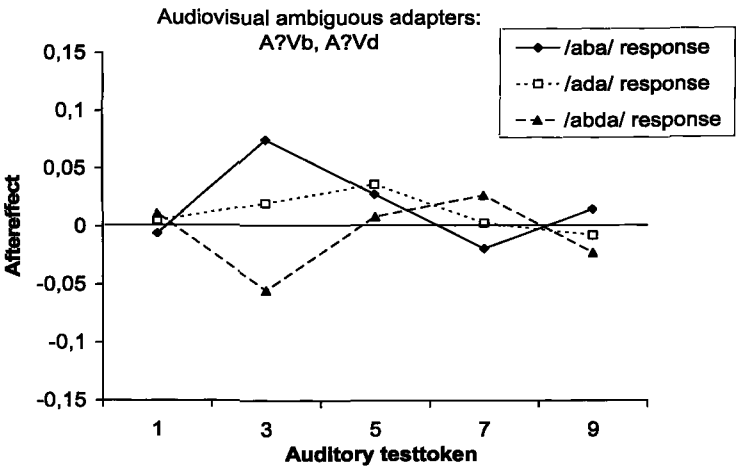


Figure 4



Figures 2, 3 and 4: Aftereffects produced by audiovisual congruent, incongruent and ambiguous exposure stimuli in Experiment 2. After-effects reflect for each auditory testtoken and each response-option separately the difference in response proportion between baseline and post testing. A positive aftereffect for a particular response-option indicates an increase in the frequency with which that response-option was chosen during posttests.

General Discussion

The main purpose of the present study was to demonstrate the existence of dissociation between aftereffects induced by audiovisual ambiguous and non-ambiguous adapter sounds when combined with congruent or incongruent visual speech.
In the original study by Bertelson et al. (Bertelson et al., 2003) it was found that exposure to audiovisual congruent and audiovisual ambiguous stimuli led to aftereffects in opposite directions. This difference, which was presented as demonstrating the dissociation between visual recalibration and selective speech adaptation, has been replicated here. On the other hand, the present experiment was run both with 8 exposures per adaptation block, as in the original study, and with 50 exposures per block. This extension produced three important findings. First, the two opposite directions of adaptation were obtained with both exposure lengths, showing that the dissociation is a robust one. Second, the aftereffects of exposure to the congruent audiovisual adapters (indicative of selective adaptation) increased significantly with number of exposures, while those of exposure with the ambiguous auditory token remained at the same level. Thus, the two basic adaptation

phenomena follow different build-up courses, providing a new dissociation, beside the one based on adaptation directions (Bertelson et al., 2003).

Third, aftereffects of adapters with non-ambiguous sounds were immune to whether visual information was congruent, incongruent or absent. This occurred despite the fact that visual information affected the way the adapting stimulus was perceived. Yet, in all three cases, equal amounts of selective speech adaptation were obtained. In addition, visual-only information did not cause any aftereffects, thus confirming previous reports that selective speech adaptation is solely determined by the acoustic dimensions of the adapting stimulus, and not the visual information or its phonetic percept when the acoustic speech signal is clear (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). The elegance of the present study is that we not only confirm these findings, but here we also demonstrate that the absence of visual effects in selective speech adaptation cannot be attributed to the idea that visual information as such is impotent, because visual information had a substantial impact on aftereffects when the sound was ambiguous. Visual information thus can affect aftereffects, but only if combined with an ambiguous sound.

Our findings are also relevant to the question why McGurk-like stimuli induced only with no trace of recalibration. If recalibration were solely driven by audiovisual discordance, one could have expected more negative aftereffects with audiovisual incongruent stimuli than the audiovisual congruent ones, because recalibration and selective speech adaptation might add up in the former case. Yet, no such effect was found here, nor was it found in previous reports (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). One way to account for is to appeal to the rule of 'inverse effectiveness'. The rule of inverse effectiveness is a general rule of information integration which states that the impact of one source of information on the other depends on the relative weights of the two (Ernst & Banks, 2002; Stein & Meredith, 1993). It not only applies to tasks that measure immediate visual bias of auditory speech (Massaro, 1987), but it has also been demonstrated in cases where aftereffects are measured like visual-haptic cue weighting (Ernst, Banks, & Bülthoff, 2000). In the present case, one expects the impact of the visual stimulus to be bigger when combined with an ambiguous sound than when combined with a non-ambiguous one. This difference in relative dominance may also determine the extent of recalibration. On this account, recalibration of speech is driven by discordance between the visual and auditory information to the extent that the auditory information is ambiguous relative to the visual one. If the sound is non-ambiguous, is cannot be recalibrated by incongruent visual distracters.

Chapter 3

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

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.

Introduction

The question of how sensory modalities cooperate in forming a coherent representation of the environment is the focus of much current work at both the behavioural and the neuroscientific level. A substantial part of that work is carried out with conflict situations, in which incongruent information about potentially the same distal event is presented to different modalities (Bertelson & De Gelder, 2003). Exposure to such conflicting inputs produces two main effects: immediate biases and aftereffects. Immediate biases are changes in the perception of stimuli in a target modality produced by the presentation of incongruent stimuli in a distracting modality. One well-known example is the ventriloquist effect, in which the perceived location of target sounds is displaced toward light flashes delivered simultaneously at some distance, in spite of instructions to ignore the lights (Bertelson, 1999; Vroomen & de Gelder, 2004a). Aftereffects are shifts following exposure to an intersensory conflict, when data in one or in both modalities are later presented alone. For the ventriloquism situation, aftereffects have been reported in which unimodal auditory localization is displaced in the direction of the light as seen in the exposure phase (Frissen et al., 2003; Radeau & Bertelson, 1974). The occurrence of such aftereffects implies that exposure to conflicting inputs recalibrates processing in the respective modalities as the previously experienced conflict is reduced.

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

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

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

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

Method

Participants

24 naïve first-year psychology students participated in the experiment.

Materials

A 9-point /aba/-/ada/ speech continuum was created and dubbed onto the video of a face articulating /aba/ or /ada/. Stimulus preparation started with a digital audio (Philips DAT-recorder) and video (Sony PCR-PC2E MiniDV) recording of a male speaker producing multiple repetitions of /aba/ and /ada/ utterances. Clearly spoken /aba/ and /ada/ tokens were selected and served as reference for the creation of the continuum. The stimuli were synthesized with the Praat program (http://www.praat.org/) (Boersma and

Weenink, 1999). The glottal excitation source used in the synthesis was estimated from a natural /aba/ by employing the inverse filtering algorithm implemented in Praat. The stimuli were 640 ms in duration with a stop consonant closure of 240 ms. A place-of-articulation continuum was created by varying the frequency of the second (F2) formant in equal steps of 39 Mel. The onset (before the closure) and offset (after the closure) frequency of the F2 was 1250 Hz. The target frequency was 1100 Hz for /aba/ and 1678 Hz for /ada. The F1 transition changed from 750 Hz to 350 Hz before the closure for both stimuli. After the closure a mirror image of the transition was used. The duration of the transition was 40 ms both before and after the closure of the consonant. The third (F3), fourth (F4), and fifth (F5) had fixed frequencies of 2500 Hz, 3200 Hz, and 4200 Hz, respectively. The amplitude and the fundamental frequency contour followed those of the original /aba/ token.

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

Procedure

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

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

In the calibration phase, the participant was presented all stimuli of the continuum in random order and categorized them as /aba/ or /ada/. Tokens from the middle of the continuum were presented more often than tokens at the extreme (6, 8, 14, 14, 14, 14, 14, 8 and 6 presentations for each of the nine stimuli, respectively). Participants were instructed to listen to each stimulus and to respond by pressing a 'b' or a 'd' on a keyboard upon hearing /aba/ or /ada/, respectively. The stimulus nearest to the

50% cross-over point was estimated via probit analysis, and this stimulus (A?) served as the most ambiguous stimulus in subsequent testing.

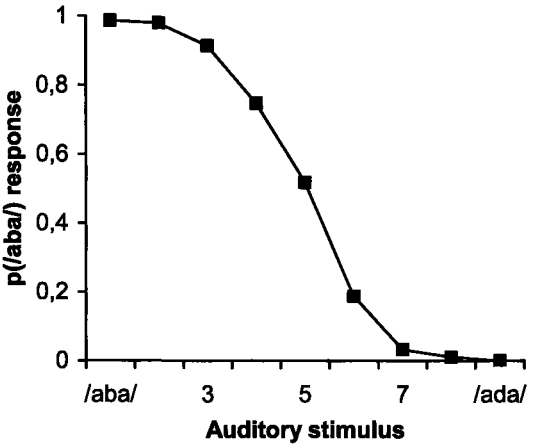
The pre-test consisted of 60 auditory-only test trials (2.5 s. ITI), divided into 20 triplets. Each triplet contained the three auditory test stimuli nearest to the individually determined phoneme boundary (A?-1, A?, A?+1). Trials within a triplet were presented in different random orders. Participants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

For the audio-visual exposure phase, participants were randomly assigned to one of four groups (6 participants each). Participants were exposed to either AbVb, AdVd, A?Vb or A?Vd for three blocks of 50 trials each (1.5 s. ITI). Five catch trials were interspersed during audio-visual exposure to ensure that participants were attending the face. Catch trials consisted of the presentation of a small white spot (12 pixels) between the lips and the nose of the speaker for three video frames (~100 ms). Participants had to press a key whenever a catch trial occurred (thus no phonetic categorization was required during the audio-visual exposure phase). Each of the three audio-visual exposure blocks was followed by an auditory-only identification task. These post-tests were the same as the pre-test, and thus consisted of 20 triplets of the three boundary stimuli (A?-1; A?; A?+1). Three quasi-random orders were used for the post-tests so that each of the three test-stimuli appeared once at each serial position.

Results

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

Figure 2



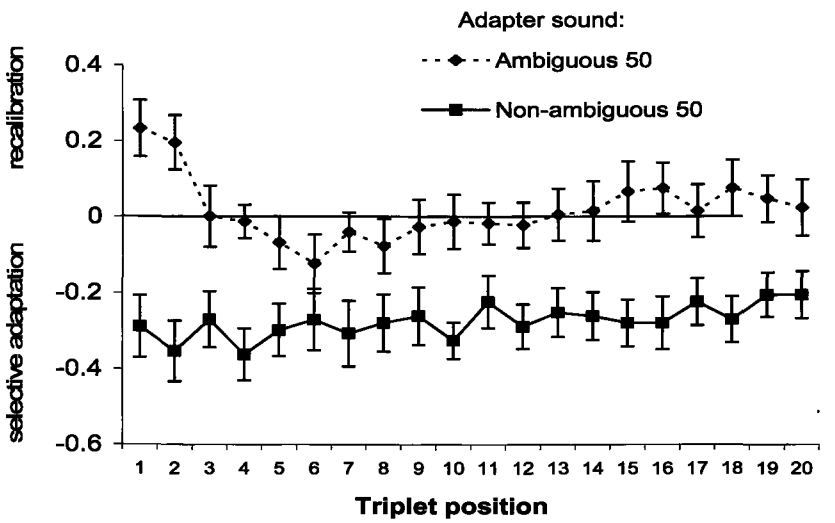
Mean proportion of /aba/ judgements for each item of the continuum in the calibration phase.

Audio-visual exposure: Participants detected on the average 91% of the catch trials, indicating that they were indeed looking at the video during exposure. Aftereffects were calculated by subtracting the proportion of /aba/ responses in the pre-test from their proportion in the post-tests, so that a positive sign referred to an increase in responses consistent with the visual distracter as seen during the exposure phase. For example, when a participant responded in the pre-test on 50% of the trials /aba/, and following exposure to A?Vb, it was 60% in the post-test, then the aftereffect was +10%.
Figure 3 shows the thus determined aftereffects as a function of the serial position of the test triplet. As in Bertelson et al., (2003), exposure to ambiguous sounds increased the number of post exposure judgments consistent with the visual distracter (i.e., more /aba/-responses after exposure to A?Vb, and more /ada/-responses after exposure to A?Vd; i.e., recalibration). The opposite effect was found after exposure to non-ambiguous sounds (fewer /aba/-responses after exposure to AbVb, and fewer /ada/-responses after exposure to AdVd; i.e., selective adaptation). As is apparent in Figure 3, the recalibration effect was short-lived and lasted for only 6 test trials (the first and second triplet positions of the test items), whereas selective adaptation lasted for the whole test.

A 2 (non-ambiguous-sound exposure vs. ambiguous-sound exposure) x 20 (triplet position) ANOVA (with the sign of the effect reversed for non-ambiguous sound exposure) on the aftereffects showed that the size of the aftereffect following exposure to non-

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

Figure 3



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

Discussion

Exposure to a particular visual speech token combined with the corresponding non-ambiguous auditory token resulted in a reduced tendency to produce that token, i.e. the typical selective speech adaptation effect. The same visual token combined instead with an ambiguous auditory token resulted in the opposite shift, a more frequent production of the response consistent with the visual adapter, indicative of crossmodal

recalibration. Thus, as in Bertelson et al., (2003), a dissociation between the two adaptation effects was obtained under otherwise identical conditions, just by manipulating the ambiguity of the auditory speech presented during adaptation.
The new finding in the present experiment is that the two effects dissipate at different rates. Whereas recalibration lasted, in the present set-up, only about 6 test trials, selective adaptation could be observed even after 60 test trials. This difference confirms that the two adaptation phenomena result from different underlying mechanisms.

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

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

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

might also be the case that ceiling effects were at play with the AbVg adapter, such that recalibration effects were overwhelmed by selective adaptation.

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

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

Conclusions

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

Chapter 4

Selective adaptation and recalibration of auditory speech perception by lipread information: The effect of exposure duration on dissipation rate.

Introduction

Exposure to audio-visual speech can modify auditory speech identification through both visual recalibration and unimodal selective adaptation. The repeated exposure to audiovisual congruent speech information causes selective adaptation: a decrease in the frequency with which that speech token is reported on a subsequent phoneme categorization task. Exposure to audio-visual speech, containing auditory ambiguous speech information produces recalibration which is observable as a post-exposure increase in visual consistent responses on the subsequent auditory-only phoneme categorization task. Exposure to audiovisual congruent and audiovisual ambiguous stimuli thus produce aftereffects in opposite directions. In addition, we observed the aftereffects to also dissipate at different rates (Vroomen et. al, 2004), providing further evidence for the suggestion that these aftereffects result from different speech processes. Whereas recalibration dissipated fast with prolonged posttesting, selective adaptation remained stable over time.

In addition to the difference in rate of dissipation, the stability of the aftereffects may also be affected differently by exposure duration, which would further dissociate the aftereffects. Here, posttests were preceded by either 8 or 32 adapter presentations in order to investigate for an effect of exposure duration on the dissipation-functions of selective adaptation and recalibration. Since aftereffects produced by the audiovisual ambiguous adapters had already dissipated after 6 testtokens in our previous dissipation study, posttesting was shortened to twelve auditory identification trials in order to reduce the total testing time.

Method

Participants

Twenty participants, all first-year students of the Tilburg University, participated in the experiment.

Materials and procedure

Stimuli, materials and procedures were the same as those adopted in the experiment presented in chapter 3, except that here aftereffects were measured in a within-subjects design and the number of adapter presentations and posttest tokens were different. Testing for an effect of exposure duration on the dissipation rates of recalibration and selective adaptation was organized in two experimental blocks. In one of the blocks the audiovisual adapters were presented 8 times successively followed by 12 auditory-only posttest trials. In the other block the audiovisual adapter was presented 32 times, also followed by 12 auditory-only posttest trials. Participants engaged in all four conditions, two

audiovisual non-ambiguous sound conditions (AbVb and AdVd) and two audiovisual ambiguous sound conditions, (A?Vb and A?Vd). Presentation order of the experimental blocks was counterbalanced over participants. Posttests were composed of four triplets of the three auditory tokens nearest to the individual phoneme boundary. Participants pressed 'b' or 'd' on a keyboard upon hearing /aba/ or /ada/. Attention was kept on the face by having the participants monitor it for occasional catch trials.

Results

Participants detected on the average 98% of the catch trials, indicating that they were indeed looking at the video during exposure. The individually determined most ambiguous auditory stimulus (A?) ranged between stimulus 4 and 6.

Aftereffects were calculated by subtracting the proportion of /aba/ responses in the pre-test from their proportion in the post-tests, such that a positive sign referred to an increase in responses consistent with the visual distracter as seen during the exposure phase

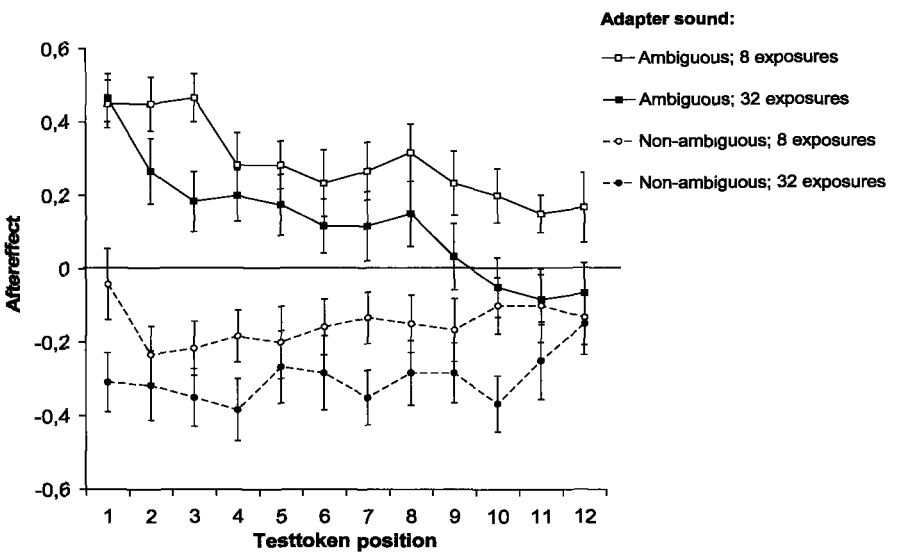
Aftereffects are displayed in figure 1. As observed in the previous experiments, the proportion exposure-consistent responses decreased after exposure to the audiovisual non-ambiguous stimuli. Audio-visual ambiguous exposure increased exposure consistent identification on the subsequent post-tests. In addition, aftereffects for audiovisual non-ambiguous exposure were increased in magnitude when adapter preceded posttests. For the audiovisual ambiguous adapters the reverse effect is observed, here, aftereffects are bigger after 8 then after 32 adapter exposures. Furthermore, as in the study presented in chapter 3, aftereffects caused by audiovisual ambiguous adapters dissipate faster than the aftereffects induced by audiovisual non-ambiguous exposure.

These observations were statistically confirmed. In a 2 (Audio-Visual non-ambiguous vs. Audio-visual ambiguous exposure) x 2 (Exposure Duration) x 12 (Testtoken Position) MANOVA (with the sign of the effect reversed for audio-visual non-ambiguous), no effect of adapter-type was found $F(1,19) < 1$, as overall, the aftereffects induced by audiovisual non-ambiguous and audiovisual ambiguous stimuli were equal in magnitude. The main effect for Testtoken Position, $F(11, 209) = 5.93$, $p < .001$, and its interaction with Adapter type, $F(11, 209) = 3.61$, $p < .001$ were significant, as the aftereffects caused by audiovisual congruent exposure remained stable whereas the aftereffect of audiovisual ambiguous exposure decreased with the presentation of more testtokens. The interaction between Adapter type and Exposure Duration was also significant, $F(1,19) = 9.71$, $p < .01$, corresponding to the reverse effect of Exposure Duration for audiovisual non-ambiguous and audiovisual ambiguous exposure.

To investigate the individual dissipation rates, two separate 2 (8 vs. 32 Exposures) x 12 (Serial Position of the Testtoken) GLM's were run for selective adaptation

and recalibration. For selective adaptation, only the Number of Exposures was significant, as more exposures increased the effect $F(1,19) = 5.55$, $p < .03$. Neither the effect of Serial Position of the Testtoken, nor the interaction with Exposure Duration was significant, (both F- values < 1). For recalibration we also observed an effect for Exposure Duration, as the effect was bigger after 8 than after 32 adapter presentations, $F(1,19) = 4.78$, $p < .05$. The effect of Serial Position of the Testtoken was also significant $F(1,19) = 48.49$, $p < .001$, as the effect dissipated in a linear trend. No interaction was found for Exposure Duration and Testtoken Position, as after 8 and 32 adapter presentations the effect dissipated in a similar trend $F(1,19) = 2.42$, $p = .14$.

Figure 1



Aftereffects as a function of the serial position in the posttest. After exposure to ambiguous sounds (A?Vb and A?Vd), the number of responses consistent with the video increased. The aftereffects produced by the adapters with ambiguous sounds were bigger after 8 than after 32 adapter repetitions. After exposure to non-ambiguous sounds (AbVb and AdVd) the number of responses consistent with the video decreased (= selective adaptation). The aftereffects produced by the adapters with clear sounds were bigger after 32 than after 8 adapter repetitions.

Discussion.

Again, exposure to audiovisual non-ambiguous sound stimuli produced selective adaptation to speech and exposure to audiovisual ambiguous sound adapters produced recalibration. Selective adaptation increased when with the presentation of more exposure stimuli and the effect did not dissipate with prolonged posttesting. The new finding here is that the number of adapter presentations did not affect the rate of dissipation for selective adaptation. In fact, no dissipation was observed for selective adaptation. The stability of selective adaptation is thus dependant neither on the size of the effect nor on the duration of exposure.

Unexpectedly, the magnitude of recalibration was actually smaller after 8 than after 32 exposure repetitions. The duration of exposure duration did not affect the rate of dissipation for recalibration. Again we observed recalibration to dissipate with the presentation of more posttest tokens, and the effect dissipated equally fast after 8 or 32 adapter presentations.

When considering the time course of the other crossmodal recalibration effects like the ventriloquist aftereffect and prism adaptation aftereffects, one would expect to find bigger recalibration-effects when exposure is prolonged (Epstein, 1975). For example, in the prism adaptation study by Fernandez-Ruiz and Diaz (1999), participants had to throw 0, 3, 9, 13 or 25 balls toward a target while wearing prisms which shifted the visual plane horizontally. Then, prisms were removed and subjects threw another 25 balls to establish aftereffects. Results clearly indicated that the magnitude of the aftereffect increased with an increasing number of throws during the exposure phase. Frissen (2005), also reported an increase in the magnitude of the ventriloquist-aftereffect when exposure was pronged. In his study participants were exposed for a period of 1, 3 or 5 minutes to a ventriloquist situation in which a sound stimulus and a light flash were presented temporally simultaneous, but at different spatial locations. The aftereffect of exposure to such a situation is observable as a shift in post-exposure sound localization towards the direction of the light-flash. The ventriloquist aftereffect was bigger in magnitude when exposure was prolonged. The number of exposure stimuli did however not affect the rate of dissipation of aftereffects. In fact, no dissipation was observed in that experiment as the magnitude of the ventriloquist aftereffect remained stable over the 90 seconds of post testing.

The study presented in chapter 5 provides further insights in the acquisition functions of recalibration and selective adaptation in speech perception.

Chapter 5

Visual Recalibration and Selective Adaptation in Auditory-Visual Speech Perception: Contrasting Build-up Courses.

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,* 572-577.

Introduction

The question of how sensory modalities cooperate in forming a coherent representation of the environment is the focus of much current research. The major part of that work is carried out with conflict situations, in which incongruent information about potentially the same distal event is presented to different modalities (see reviews by Bertelson & de Gelder, 2004; de Gelder & Bertelson, 2003).

Exposure to such conflicting inputs produces two main effects: immediate biases and after-effects. By immediate biases are meant effects of incongruent inputs in a distracting modality on the perception of corresponding inputs in a target modality. For example, in the so-called ventriloquist illusion, the perceived location of target sounds is displaced toward light flashes delivered simultaneously at some distance, in spite of instructions to ignore the latter (Bertelson, 1999). After-effects (henceforth "AEs") are shifts in perception observed following exposure to an inter-modal conflict, when data in one or in both modalities are later presented alone. For the ventriloquism situation, unimodal sound localization responses are, after exposure to synchronized but spatially discordant sound bursts and light flashes, shifted in the direction of the distracting flashes (Radeau & Bertelson, 1974; Recanzone, 1998). The occurrence of AEs has generally been taken as implying that exposure to incongruence between corresponding inputs in different modalities recalibrates processing in one or both modalities in a way that eliminates (or at least reduces) the perceived discordance. Although immediate biases and recalibration have consistently been demonstrated for spatial conflict situations, the evidence has long been less complete for conflicts regarding event identities. Here, biases were often reported, but, for some time, no recalibration. The main example is the conflict resulting from the acoustic delivery of a particular speech utterance in synchrony with the optical presentation of a face articulating a visually incongruent utterance. As originally reported by McGurk and MacDonald (1976), this kind of situation generally produces strong immediate biases of the auditory percept towards the speechread distracter, a phenomenon now generally called "the McGurk effect". For instance, auditory /ba/ combined with visual /ga/ is often heard as /da/. On the other hand, no demonstration of AEs consequent upon exposure to McGurk situations had until recently been reported, and results in the literature (Roberts & Summerfield 1981; Saldaña & Rosenblum, 1994) were taken as implying that such exposure produces no recalibration, possibly revealing a basic difference between identity and spatial conflicts (Rosenblum, 1994).

Using a new type of adapting situation, we have however now succeeded in demonstrating the latter kind of recalibration (Bertelson, Vroomen, & de Gelder, 2003). Our exposure situation involved bimodal stimulus pairs in which the auditory component was each participant's most ambiguous speech utterance from an /aba/ - /ada/ continuum (A?), and the visual component featured the articulation of either of the two end points,

/aba/ or /ada/. Following the habitual conflict adaptation paradigm, auditory identification tests, using the ambiguous utterance and two slightly less ambiguous ones as material, were administered after exposure to bimodal adapters with either the /aba/ or the /ada/ visual component. As expected, /aba/ responses were more frequent after exposure with visual /aba/ than with visual /ada/, thus revealing recalibration.

Our reason for using an ambiguous auditory adapter was to avoid the occurrence of the so-called selective speech adaptation phenomenon, in which repeated exposure to a non-ambiguous auditory speech utterance causes a reduction in the frequency with which that utterance is reported on subsequent identification trials (Eimas & Corbit, 1973; Samuel, 1986). Selective speech adaptation is thus, like recalibration, an adaptation phenomenon that manifests itself by AEs but, unlike recalibration, does not depend on the co-occurrence of conflicting inputs in another modality. If our bimodal exposure had been run with unambiguous auditory utterances, e.g. auditory /aba/ paired with visual /ada/, the same outcome on post-test, more /ada/ responses, could have been equally attributed to selective speech adaptation from auditory /aba/ as to recalibration of speech identification by the visual distracter /ada/.

That exposure to bimodal pairs with unambiguous auditory speech utterances from our material can actually produce selective speech adaptation was demonstrated in the same study (Bertelson et al., 2003, Exp. 2) by exposing participants to congruent and unambiguous audio-visual pairs, either auditory /aba/ combined with visual /aba/, or auditory /ada/ combined with visual /ada/. In this new condition, exposure effectively resulted in a reduction of the proportion of responses consistent with the bimodal adapter. Fewer /aba/ responses occurred after exposure to bimodal /aba/ than to bimodal /ada/, the outcome opposite the one obtained when the same visual /aba/ was paired with the ambiguous auditory utterance. The congruent visual component presumably played no role in the causation of selective adaptation, but its presence made each congruent non-ambiguous adapting pair undistinguishable from the pair with the same visual component and the ambiguous auditory component, as was shown in separate identification tests.

Additional evidence for the dissociation between two adaptation phenomena was provided more recently in a study showing that they dissipate following different courses (Vroomen, van Linden, Keetels, de Gelder & Bertelson, 2004). The present study is focuses on the build-up of the AEs successive presentations of the bimodal adapters of our original study (Bertelson et al., 2003). Two of these, making up the ambiguous sound condition, consisted of the participant's most ambiguous auditory utterance A?, paired across successive presentations either with visual /aba/ (pair A?Vb) or with visual /ada/ (pair A?Vd). The other two adapters, making up the non-ambiguous sounds condition, consisted of auditory /aba/ paired with visual /aba/ (pair AbVb) and of auditory /ada/ paired with visual /ada/ (pair AdVd). Following the earlier findings, the ambiguous sound condition was

expected to produce no selective speech adaptation, because of the ambiguity of the auditory component, but to cause recalibration in the direction of the incongruent visual component. In contrast, the non-ambiguous sounds condition was expected to produce selective adaptation, because of the non-ambiguous quality of each auditory component, but no recalibration, because of the absence of phonetic incongruence between auditory and visual components. The adapters were presented in continuous series of trials, and auditory AEs were measured at several successive points during each series. A first group of participants was tested with adaptation blocks running to 64 trials. Their results revealed an unexpected reversal in the build-up course of adaptation in the ambiguous sound condition. To check on this finding, the number of exposure trials was extended to 256 for a second group of participants.

Method

Participants

Two groups of 25 students from Tilburg University participated in one experimental session. Those in Group 64 were administered 64 trials long exposure blocks, and those in Group 256, 256 trials long blocks.

Materials

Details of the stimuli have been provided in an earlier paper (Vroomen et al., 2004). In short, a 9-point /aba/-/ada/ speech continuum was created by varying the frequency of the second (F2) formant in equal steps. The end-point auditory utterances and the individually determined most ambiguous one were dubbed onto the video of a face that articulated /aba/ or /ada/.

Procedure

For both groups, the session involved three successive phases: calibration, then pretest, followed by a bimodal audio-visual exposure phase, interspersed with post-test trials. The calibration phase served to determine, for each participant individually, the sound on the continuum that was nearest to her/his /aba/-/ada/ phoneme boundary. It consisted of 98 trials in which each of the nine sounds was presented in random order at 1.5 s inter-trials intervals. Sounds from the middle of the continuum were presented more often than those from the extremes (6, 8, 14, 14, 14, 14, 14, 8 and 6 presentations for each of the nine items, respectively). The participant classified the sound as /aba/ or /ada/ by pressing one of two keys. The participant's 50% cross-over point was estimated via probit analysis, and the continuum item nearest to that point (A?) served as auditory component in the bimodal exposure trials of the ambiguous-sound condition.

In the pretest phase, the participant gave dichotomous key-pressing classification responses to her/his most ambiguous sound A?, as well as to its two immediate continuum neighbors (A?-1 and A?+1). These three test sounds were presented in balanced order across 20 successive triplets. The 60 presentations followed each other without interruption at 2.5 s ITIs.

The audio-visual exposure phase consisted of eight adaptation blocks, two for each of the four bimodal adapters AbVb, AdVd, A?Vb and A?Vd. For Group 64, the adapters were presented 64 times in each block at 1.5 s ITIs, and two triplets of auditory identification post-tests, identical to those in the pretest phase, were interpolated after 1, 2, 4, 8, 16, 32, and 64 exposures. For Group 256, there were 256 exposures per block presented at 0.85 s ITIs, and post-tests were interpolated at the same locations as for Group 64, plus locations 128 and 256. No phonetic decisions were required during the audio-visual exposure phase, but participants had to press a special key on every presentation of a visual catch stimulus (a 12 pixels white spot flashed for 100 ms between the nose and the upper lip of the talker). Five such catch trials were interpolated at random moments during each block, in order to ensure attention to the face. Presentations of the different types of blocks were counterbalanced across participants.

Results

The individually determined most ambiguous auditory stimuli (A?) ranged between utterances 4 and 6. During bimodal exposure, participants detected 93% of the visual catch stimuli, indicating that they were effectively attending to the face. AEs were calculated (as in Bertelson et al., 2003), by the difference in the proportion of /aba/ responses obtained after exposure to respectively A?Vb and A?Vd (ambiguous sound condition) or after AbVb and AdVd (non-ambiguous sound condition). Recalibration thus manifests itself by positive AEs and selective adaptation by negative ones. Mean AEs as functions of number of exposure trials across each type of block are shown in Figure 1.

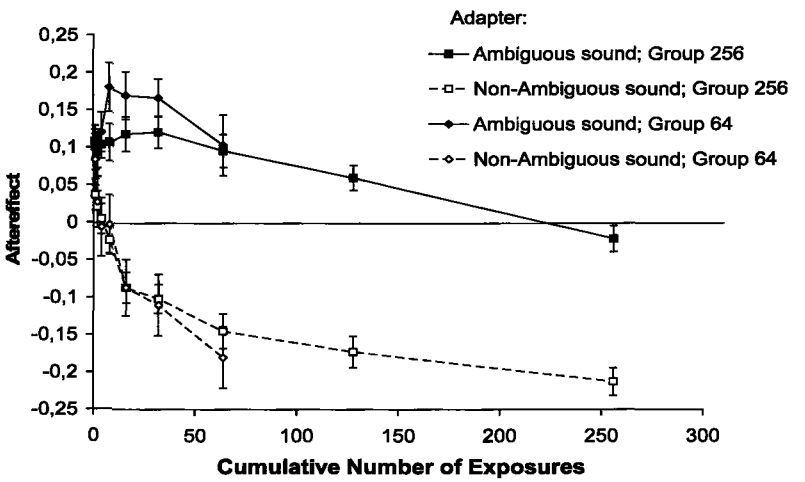
As a first step, the data from each group were submitted to two separate two-factor (auditory ambiguity and number of exposures) MANOVAs. For Group 64, both main effects, auditory ambiguity, $F(1, 24) = 52.6$, $p < .001$, and number of exposures, $F(6, 144) = 6.37$ $p < .001$, and their interaction $F(6, 144) = 15.3$, $p < .001$, were significant. The results were identical for Group 256: condition, $F(1, 24) = 57.1$, $p < .001$, number of exposures, $F(8, 192) = 31.8$, $p <. 001$, interaction, $F(8, 192) = 13.2$, $p < .001$. The effects of auditory ambiguity correspond to the fact that AEs were mainly positive with ambiguous adapters and mainly negative with non-ambiguous ones. The interactions reflect the fact, clearly visible in Figure 1, that AEs follow different courses in the two conditions, monotonically decreasing in the non-ambiguous sound condition and going up and then down in the ambiguous sound condition.

In the ambiguous sound condition, AEs appear to peak higher in Group 64 than in Group 256. To check on the significance of that difference, the data of group 64 were entered together with those for the first 64 exposures of group 256 into a two-factor(Group and Number of Exposures) MANOVA. No significant main effect of Group, $F(1, 48) < 1$, nor any interaction with number of exposures, $F(6, 288) = 1.52$, $p > .10$, emerged. Thus, the observed difference probably resulted from random variations among participants in the two groups. A similar absence of group difference could be expected for the non-ambiguous condition on the basis of the data in Figure 1, and was confirmed by MANOVA, both $Fs < 1$.

Given the absence of significant difference, the data from the two groups could be pooled and the resulting values (shown in Figure 2) submitted to two General Linear Model (GLM) analyses, allowing the examination of trends. For the ambiguous sound condition, the analysis produced a significant quadratic component, $F(1,49) = 7.34$, $p < .01$, and the linear component, $F(1,49) = 1.65$, $p > .20$, was non-significant. The quadratic component reflects the fact that the AE rose, reached a plateau and then went down. For the non-ambiguous condition (lower part of Figure 2), GLM produced a highly significant linear component, $F(1,49) = 91.2$, $p < .001$, as well as significant quadratic, $F(1,49) = 21.6$, $p < .001$, and cubic, $F(1,49) = 11.6$, $p < .001$, ones. The linear component reflects the monotonic decreasing slope of the curve and the two higher order components, its gradual flattening. Finally, application of GLM to the 64 to 256 exposures AEs of group 256 produced significant linear trends ($p < .01$ for both conditions) and no higher order trends (all $Fs < 1$).

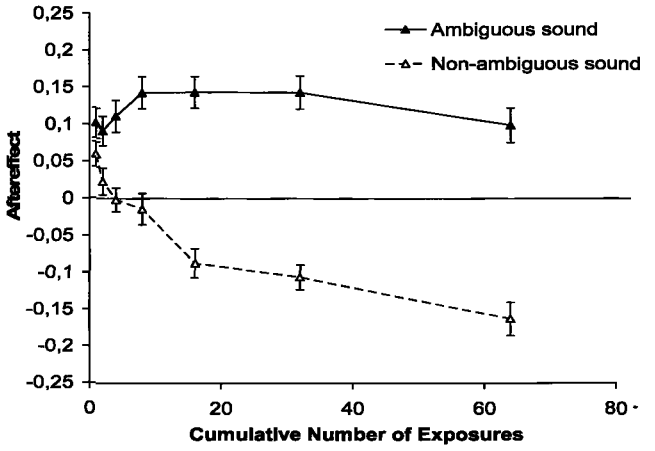
Two somewhat unexpected aspects of the build-up courses deserve attention. Both concern the starting points of the curves. In the ambiguous sound condition, a substantial positive AE , significantly superior to zero, $t(49) = 5.98$, $p < .001$, already occurs after the single first presentation of the bimodal adapter. In the non-ambiguous sound condition, a significant positive AE, $t(49) = 2.98$, $p < .005$, occurs after the first presentation. It gives way to the expected negative values on succeeding exposures. Possible reasons for these effects will be examined in the Discussion.

Figure 1



Mean after-effects as functions of cumulative number of exposures in the ambiguous sound condition (adapters A?Vb and A?Vd) and the non-ambiguous sound condition (adapters AbVb and AdVd) for Group 64 (exposures 1 to 64) and Group 256 (exposures 1 to 256). Aftereffects are the differences between the proportions /aba/ responses obtained with adapters A?Vb and A?Vd (ambiguous condition) or with adapters AbVb and AdVd (non-ambiguous condition).

Figure 2



Mean aftereffects as functions of cumulative number of exposures (1 to 64), for the pooled data of Groups 64 and 256, in the ambiguous sound condition (adapters A?Vb and A?Vd) and the non-ambiguous sound condition (adapters AbVb and AdVd).

Discussion

The present experiment examined the way the contrasting auditory AEs obtained in our earlier studies (Bertelson et al., 2003; Vroomen et al., 2004), after exposure to bimodal pairs with respectively ambiguous and non-ambiguous auditory components, build up. Two main results emerged.

First, the main directions in which AEs develop are the same as in the earlier experiments. After eight presentations (the level of exposure used throughout in the original study) AEs went in the direction of the visual distracter in the ambiguous sound conditions, and in the opposite direction (away from the congruent bimodal adapter) in the non-ambiguous sound conditions. This contrast, which was presented as demonstrating the dissociation between recalibration and selective speech adaptation, is thus replicated. The fact established in the original study (Bertelson et al., 2003) that in our material corresponding bimodal adapters differing only at the level of auditory ambiguity (like A?Vb and AbVb) are perceptually undistinguishable should be stressed again at this point. It carries the important implication that the contrasting AEs obtained in the two conditions cannot have originated in deliberate post-perceptual strategies, and must be of perceptual nature.

Second, the respective developments not only go in opposite directions, but also follow different courses, monotonically descending for non-ambiguous sounds, and curvilinear, with a rapid early build-up followed by a plateau and then a gradual decline, for ambiguous sounds.

In our initial paper (Bertelson et al., 2003) we proposed that the AEs obtained in the ambiguous sound condition reflected essentially recalibration, and those in the non-ambiguous sound condition, essentially selective adaptation. Let us now examine how the build-up results affect these proposals.

For the non-ambiguous sound condition, the monotonic descent of the curve is consistent with the interpretation in terms of cumulative selective adaptation, and the gradual deceleration of that descent suggests evolution toward some asymptotic value. The fact that the descending curve starts, after the first exposure trial, not at zero or already at some small negative level, but at a significantly positive one, may seem surprising. A possible explanation would be that presentation of a non-ambiguous (end-point) auditory utterance produces not only selective adaptation but also some priming or repetition effect, i. e. moving perception of the ambiguous test utterance in the direction of the just presented non-ambiguous one, the direction opposite that of selective adaptation. If the priming effect, in contrast to the cumulative selective adaptation, was constant from trial to trial, it might overrun the latter on early presentations but be overtaken by it later on, thus producing the pattern observed in the figure.

For the ambiguous sound condition, the main finding is the curvilinear development course. That an initial positive growth gradually gives way to a decline is supported by the quadratic trend obtained for the pooled data over exposures 1 to 64, and the reality of the final decline receives additional support from the descending linear trend obtained over the last three post-tests of group 256. What mechanism could produce such a pattern?

The ascending part of the curve in all probability reflects increasing recalibration. A question similar to the one discussed for the non-ambiguous sound data may be raised concerning the significant AE already present after the first exposure. The two cases are however not identical. In the non-ambiguous condition, the first trial AE went in the direction opposite the later selective adaptation, thus requiring a different explanation, like the one through a priming effect that we proposed. For the ambiguous condition, the first trial AE goes in the same positive direction as the later build-up, so that it can just be the effect of a very rapid, or one-trial, recalibration process. That priming would also play some role cannot be excluded on the basis of the present data, and the possibility should be a matter for future investigations.

Regarding the later decline, there is of course no apparent reason why a learning phenomenon like recalibration would reverse itself at some point. Some separate process must be involved here. The most likely possibility is a selective adaptation process running in parallel with recalibration and eventually counterbalancing it. This process could start as soon as some sufficient exposure to non-ambiguous sounding inputs has occurred. A basis for selective adaptation can be provided on each trial since pairing the ambiguous utterance with a (non-ambiguous) visual component makes it sound non-ambiguous (through the McGurk effect). Of course, the same bimodal pair also produces recalibration because of the discrepancy between the ambiguous utterance and the non-ambiguous visual component. Whether the progressive disambiguation of the ambiguous utterance consequent on recalibration results in a faster accumulation of selective adaptation is on the other hand not something the data tell us.

In conclusion, the present data impose a revision of our earlier interpretation of the adaptation observed in the ambiguous sound condition as reflecting exclusively recalibration. Selective adaptation would be implied also. The relative strengths of the two effects would change with cumulative exposure to the adapting pair, with recalibration dominating at early stages and being later on progressively counterbalanced by selective adaptation.

Bertelson et al. (2003) noted that while both conflict-based recalibration and conflict-free sensory adaptation have been demonstrated in several other perceptual domains, their interaction had rarely been considered within the same experimental situation. A relevant case has been revealed by recent work concerning the influence of

lexical context on auditory phoneme identification. In an important study, Samuel (2001) has used the selective adaptation paradigm to demonstrate such lexical influences in the absence of contamination by post-perceptual adjustments. His participants were exposed to repeated presentations of an ambiguous /s/ - /ʃ/ sound in the context of either an /s/-final word (e.g. /bronchiti?/, from bronchitis), or an /ʃ/-final one (e.g. /demoli?/, from demolish). In post-tests involving identification of the ambiguous /s/ - /ʃ/, fewer reports of a particular alternative were obtained after exposure to words favouring that alternative. Samuel concluded that the lexically induced phoneme had produced selective adaptation, in the same manner as an acoustically delivered one.

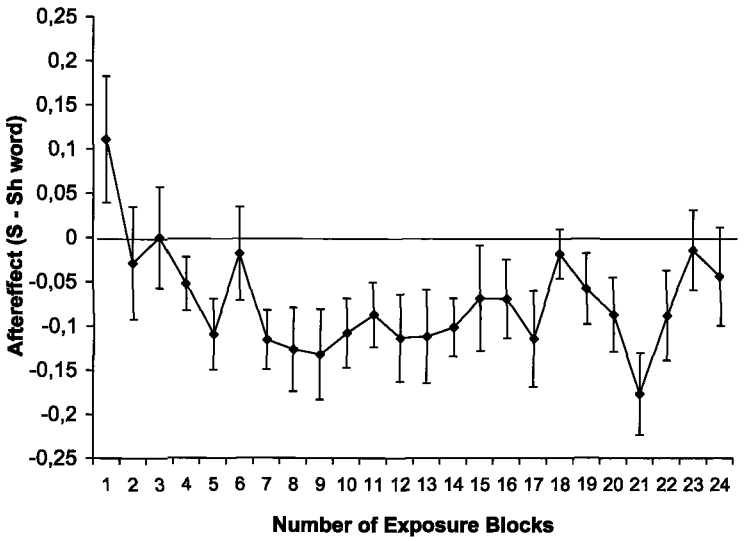
More recently, though, Norris, McQueen, and Cutler (2003) exposed listeners with similar materials, but instead of selective adaptation, they observed recalibration (or, in their terms, "perceptual learning"(1)). For instance, they replaced the final fricative (/f/ or /s/) from critical Dutch words by an ambiguous sound, intermediate between /f/ and /s/. Listeners heard this ambiguous utterance either in /f/-final words (e.g., /witlo?/, from witlof, chicory) or in /s/-final words (e.g., /naaldbo?/, from naaldbos, pine forest). Listeners who heard /?/ in /f/-final words were in subsequent testing more likely to report /f/, whereas those who heard /?/ in /s/-final words were more likely to report /s/. Thus, exposure to what seems to be very similar materials caused in one study (Samuel, 2001) selective adaptation, and the other (Norris et al., 2003) recalibration.

There are several differences between the two experiments that may explain the contradiction. For instance, Samuel used a straightforward selective adaptation method, while Norris et al. procedure involved less traditional procedures, like embedding the adapters in a larger number of neutral fillers. Our results however suggest that the critical factor may be the amount of exposure received by the participants. Norris et al. (2003) exposed their participants to just 20 inducing utterances, embedded in a single block among 180 fillers, while Samuel's (2001) procedure involved, for each of the two inducers, 24 blocks, each consisting of 32 inducers (no fillers) followed by 8 post-tests. If the lexical effects taking place in these experiments involve, like the crossmodal effects studied here, an early phase dominated by recalibration (or perceptual learning) and a later phase dominated by selective adaptation, then a short adaptation phase (as in Norris et al.) may reveal mainly recalibration, which, with the kind of massive exposure carried out by Samuel, would be overtaken by selective adaptation.

In his paper, Samuel (2001) reported only the mean adaptation effects over the whole experimental session. However, since series of posttests were carried out after each of the 24 adaptation blocks, the data contained all the necessary information concerning the build-up course. Samuel has kindly made these data available to us. Figure 3 shows AEs for successive adaptation blocks. Negative differences are observed for the clear majority of blocks posterior to block 3, showing the expected dominant role of selective

adaptation. But a positive difference, possibly indicative of recalibration, obtains on block 1 (i.e. after 32 adapter presentations), and progressively gives way to negative ones on following blocks. Thus, the succession observed in our ambiguous sound condition of a pattern dominated by recalibration and of one dominated by selective adaptation might be present in Samuel's data as well.

Figure 3



Mean aftereffects, averaged across lexical contexts, as functions of exposure blocks, in the experiment of Samuel (2001). Exposure stimuli were words with final /s/ or /ʃ/, in which the final fricative had been replaced by an ambiguous intermediate sound (e. g. /bronchiti?/, from bronchitis), or /demoli?/, from demolish). Tests in which items from an 8-step /is/ - /iʃ/continuum were categorized as /is/ or /iʃ/ were run after each block of 32 exposures. Aftereffects are measured by the proportion identifications consistent with the lexical inducers. (Data courtesy of Arthur Samuel).

Chapter 6

Audio-Visual Speech Recalibration and Selective Adaptation in Children.

van Linden, S., & Vroomen, J. (in press). Audiovisual Speech Recalibration and Selective Adaptation in Children. *Journal of Child Language, x*, xxx-xxx.

Introduction

The use of lipread information for speech perception has not only been demonstrated in adults, but also in children and infants. Several developmental studies suggest that integration of visual and auditory speech is present early in life. For example, infants at 4.5 months of age prefer to look at a face in which the articulatory gestures are in congruence with the heard vowel rather than to a face with incongruent speech gestures (Kuhl & Meltzoff, 1982). Five-month-olds are also susceptible to the McGurk-illusion as for example demonstrated by Rosenblum, Schmuckler and Johnson (1997). (see also Burnham & Dodd, 2004; Desjardins & Werker, 2004).

Despite demonstrations that lipreading contributes to speech perception in early infancy, it has also been demonstrated that the impact of lipreading to speech perception increases with age. In the original study by McGurk and MacDonald (1976), adults (18 – 40 yr) were compared with children. Results showed that adults were more influenced by incongruent visual information than two younger groups (3-4, and 7-8 years) that did not differ consistently. This developmental trend has been confirmed in later studies (Burnham & Sekiyama, 2004; Sekiyama, Burnham, Tam, & Erdener, 2003). For example, Hockley and Polka (1994) presented audiovisual conflicting stimuli to five different age-groups: five-, seven-, nine-, eleven-year-olds and adults. Results again showed that the impact of lipreading on speech identification increased with age, even beyond the age of twelve, as only half of the eleven-year-old group responded in an adult-like manner. Likewise, Massaro (1984) compared six-year-olds with adults on an audio-visual speech identification task and found that children had only half the visual influence shown by adults. In this study, children and adults also performed a secondary task in which they indicated whether or not the mouth had moved. Despite this attentional focus on the region of the mouth, the difference in visual influence between the two age groups remained, thus suggesting that it was unlikely that children's low susceptibility to visual information was due to paying less attention to the face.

In this present study, we further explored the developmental trend in the use of lipread information. So far, all previous studies have looked at the biasing effect that a lipread stimulus has on a simultaneously presented auditory speech token. Recently, though, it has been shown that besides this bias effect, lipreading may also serve a very different role in speech perception, namely in recalibration of ambiguous auditory speech (Bertelson et al., 2003; van Linden & Vroomen, in press; Vroomen, van Linden, de Gelder, & Bertelson, 2007; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004). On this view, lipread information serves as a 'teacher' of the phonetic system on how to interpret an initially ambiguous speech sound. Recalibration by lipread speech has been demonstrated by exposing participants to an auditory ambiguous speech token (henceforth

A?) intermediate between /aba/ and /ada/ dubbed onto a visible face articulating either /aba/ or /ada/ (A?Vb and A?Vd, respectively). Following a short exposure phase (usually in the order of eight trials) to either one of these audiovisual stimuli, participants were given immediately thereafter auditory-only test trials in which they categorized the auditory ambiguous token (A?) and its two closest neighbors on the phonetic continuum (A?-1 and A?+1) as either /aba/ or /ada/. Results showed that following exposure to A?Vb, participants were in the subsequently delivered test trials more likely to categorize the ambiguous test tokens as /aba/ if compared with exposure to A?Vd. The occurrence of this aftereffect (i.e., the difference in the percentage of /aba/ responses on test trials following exposure to /aba/ minus exposure to /ada/) represents the basic recalibration phenomenon. Presumably, listeners resolved during the exposure phase the conflict between the heard and the lipread speech information by shifting the phonetic boundary such that the ambiguous sound was categorized in accord with the lipread information it was combined with. This learning effect could then in turn be observed as an aftereffect in the test trials.

To control that this phenomenon was not due to some kind of priming or response bias (e.g., in test trials participants might simply repeat what they had heard during the exposure phase), participants in the study by Bertelson et al. (2003) were also exposed to auditory unambiguous and audiovisual congruent adapters AbVb and AdVd. These adapters were not expected to evoke recalibration (because there was no conflict between the heard and seen information), but selective speech adaptation (Eimas & Corbit, 1973). Unlike recalibration, selective speech adaptation does not depend upon exposure to an audiovisual conflict, but upon prolonged exposure to an auditory unambiguous speech sound as it probably reflects 'fatigue' of some of the relevant acoustic processes (Eimas & Corbit; 1973, Samuel, 1986). In line with this prediction, following exposure to AbVb, participants were in subsequently delivered test trials less likely to categorize the auditory ambiguous test tokens as /aba/ if compared with exposure to AdVd, an effect that was in the opposite direction of recalibration. Further tests also showed that participants were unable to distinguish the exposure stimuli with ambiguous speech sounds (causing recalibration) from non-ambiguous speech sounds (causing selective adaptation). Participants could thus not reliably discriminate between A?Vb and AbVb, and between A?Vd and AdVd when asked to do so in a discrimination task. This therefore excluded the possibility that results were caused by any deliberate response strategy (Bertelson et. al., 2003).

Here, we wanted to test whether there is, as in the case of the McGurk-effect, a developmental trend in the use of visual speech for the purpose of recalibration. Recalibration of auditory speech by lipread information has never been tested in children,

as it has been demonstrated only very recently in adults. In studies with adults, it has been observed that those who are more susceptible to the McGurk-illusion will also tend to show larger recalibration effects (van Linden & Vroomen, in press). If one assumes, therefore, that the contribution of lipreading to speech in general increases with age, one might expect a corresponding trend in recalibration. Younger children might on this view thus show smaller recalibration effects than older ones.

As concerns selective speech adaptation, only a few studies have addressed developmental trends (Sussman, 1993; Sussman & Carney, 1989). Sussman and Carney (1989) explored selective speech adaptation in five-to-six year-olds, seven- to-eight year-olds, nine-to-ten year-olds, and adults. They observed that when adapted to /ba/, only adults showed a boundary shift on a /ba-/da/ continuum indicative of selective speech adaptation (i.e., more /da/ responses). For the seven- to eight year-old and nine- to ten-year-old children, no boundary shift was observed, while the youngest age group actually showed a shift in the opposite direction (i.e. more responses consistent with the exposure stimuli), a finding for which no explanation was offered. In a later study, it was found that the /ba/ token was not a particularly powerful adapter, and when adapted to /da/, both five-year-olds and adults showed a boundary shift (Sussman, 1993). The differences between children and adults were explained by the children's inability to process all acoustic details, as the youngest children were also less proficient in discriminating the speech tokens.

In the present study, we explored the occurrence of both recalibration and selective speech adaptation in children. To test for cross-age differences, a group of five-year-olds and eight-year-olds participated, as in the original studies by McGurk and MacDonald (1976) and Sussman and Carney (1989). Following exposure to the auditory ambiguous and incongruent audiovisual speech tokens A?Vb and A?Vd, we expected to find aftereffects indicative of recalibration (i.e., more /aba/ responses following exposure to A?Vb than A?Vd). This effect might be smaller in the younger age group. Following exposure to the auditory non-ambiguous and congruent audiovisual speech tokens AbVb and AdVd, we expected to find aftereffects indicative of selective speech adaptation (i.e., less /aba/ responses following exposure to AbVb than AdVd), and this effect might again be smaller in the younger age group.

It should be noted on beforehand, though, that the magnitude of selective speech adaptation was expected to be smaller than that of recalibration for two reasons. First, as demonstrated by Sussman and Carney (1989), selective speech adaptation effects may be absent or even be reversed in five-to-six year-olds. Second, the optimal number of exposure trials for demonstrating recalibration is smaller than for demonstrating selective speech adaptation. Previous studies have shown that recalibration occurs fast and is

clearly visible after only eight exposure trials. In contrast, selective speech adaptation requires more exposure trials as it increases linearly with number of exposures (Vroomen et al., 2007). Given that the focus of the current study was on recalibration, we used only a limited amount of exposure trials (eight) that was optimal for obtaining recalibration, but not selective speech adaptation.

However, despite the fact that adapters with non-ambiguous speech sounds AbVb and AdVd were not expected to produce big selective speech adaptation effects, we included them in the design of our study because they allowed us to address a methodological concern. One possible strategy that children may adopt when tested in an 'exposure - test' design is that, whenever unsure, they repeat the phoneme heard during the exposure phase. Such a deliberate response strategy would effectively mask the manifestation of selective speech adaptation while it would boost, by equal amounts, the manifestation of recalibration. This problem might become particularly pertinent when comparing across age groups, as different-aged children might adopt different strategies. This strategy might also explain why children sometimes display no selective speech adaptation effects, and why five-to-six year-olds may actually show a reversed trend (Sussman & Carney, 1989).

In the present study, we sought to alleviate this response bias problem by comparing the aftereffects of audiovisual adapters containing ambiguous sounds with those containing non-ambiguous sounds (i.e., we subtracted (A?Vb – A?Vd) - (AbVb – AdVd)). This comparison between aftereffects of ambiguous versus non-ambiguous sounds annuls any contribution of deliberate response strategies to the extent that this contribution is equally big in both cases. This, though, can safely be assumed to be the case because even adults have great difficulty discriminating adapter stimuli containing ambiguous versus non-ambiguous sounds when asked to do so (Bertelson et. al., 2003). Hence, one can safely predict that if recalibration is at stake, aftereffects of audiovisual adapters containing ambiguous sounds will be more positive than those containing non-ambiguous sounds independent of participants' response biases.

Method

Participants

Twenty five-year-olds and twenty eight-year-olds from a public elementary school in a Dutch village, Hapert, were tested. All children were native Dutch speakers, and only children with normal age-related speaking and hearing proficiency, as judged by their teacher, participated in the experiment. Eight boys and twelve girls with mean age of five years and five months, and ten boys and ten girls with a mean age of eight years and six months were tested. Data from two of the five-year-olds were discarded from analyses

because one child completed only two experimental sessions, while the other frequently failed to respond during the test. This left eighteen children in the younger age group.

Stimuli

Details of the stimuli have been described elsewhere (Bertelson et. al., 2003). In short, a 9-point /aba/-/ada/ speech continuum was created by varying the frequency of the second (F2) formant in equal steps. The midpoint token of this continuum (A?), the most ambiguous one, was used to create the audio-visual ambiguous exposure stimuli. It was dubbed onto the video of a face articulating /aba/ or /ada/, resulting in the two exposure stimuli used to induce recalibration: A?Vb and A?Vd. The two endpoint tokens, clear /aba/ and clear /ada/ were dubbed onto the congruent lip gestures to create the exposure stimuli AbVb and AdVd for inducing selective adaptation. The three tokens closest to the phoneme boundary, A?, and its two neighbors on the continuum (A?-1 and A?+1), were used for the test trials.

Procedure

The children were tested individually in a quiet room in their school. They were seated in front of a table at a distance of about 50 cm from a 15-inch laptop (Dell Inspiron 2650) on which the visual stimuli were presented. Sounds were delivered over two loudspeakers that were placed besides the screen. The experimenter sat at the side of the table in front of a keyboard. Testing began with a series of practice trails followed by four experimental sessions. Total testing took about half an hour with small pauses in between sessions. Testing was completed in a single day.

There were four experimental sessions, each consisting of four exposure–test blocks. During an exposure-test block, one of the four bimodal exposure stimuli (AbVb, AdVd, A?Vb or A?Vd) was successively presented eight times (ISI = 1 s.), followed by six auditory-only test tokens. The test tokens consisted of the three midpoint tokens of the continuum (A?-1, A? and A? +1), presented twice in counterbalanced order. All four bimodal exposure stimuli were presented once in an experimental session. Total testing thus consisted of 16 exposure-test blocks (four sessions for each of the four exposure stimuli). Presentation order of the exposure stimuli was counterbalanced over the experimental sessions. Participants were asked to silently watch the video during exposure. In the test phase, the video turned grey. After participants heard a test stimulus, they were required to report orally whether they had heard /aba/ or /ada/. The experimenter then pressed a corresponding key on the keyboard. Trials in which the child was inattentive or failed to respond were discarded from analyses. The next test trial

started after a key press by the experimenter. A practice session containing each of the four different exposure test blocks was included at the start of the experiment. Practice continued until the child completely understood the task and successfully completed at least two exposure test blocks. When necessary, the child was encouraged to keep attention on the face during audio-visual exposure.

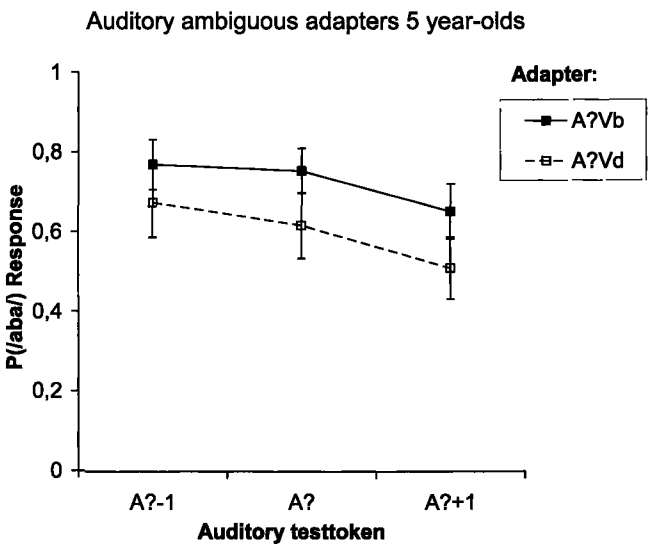
Results

Figures 1 through 4 present performance in the test phase for each exposure condition and age group. Figure 1 and 2 show the results for the five years-olds, and figure 3 and 4 for the eight years-olds; figures 1 and 3 show performance following exposure to the ambiguous adapters A?Vb and A?Vd, and figures 2 and 4 for the non-ambiguous adapters AbVB and AdVd. Each graph shows the proportion of /aba/ responses for the ambiguous token (A?), and its two neighbors on the phonetic continuum (A?-1, or A?+1). The first thing to note is that for both the non-ambiguous and ambiguous exposure conditions, children responded more often /aba/ following exposure to 'aba'-adapters (AbVb and A?Vb) than following exposure to 'ada'-adapters (AdVd and A?Vd). As already mentioned, this response tendency to repeat the phoneme heard during the exposure phase will increase the aftereffect thought to reflect recalibration (i.e., more responses consistent with the adapter) and reduce, by equal amounts, the aftereffect thought to reflect selective speech adaptation (i.e., less responses consistent with the adapter). We took precautions for this response tendency, because our design allowed us to subtract out this effect by subtracting aftereffects of non-ambiguous sounds (AbVb vs. AdVd) from ambiguous sounds (A?Vb vs. A?Vd). Recalibration devoid of response bias then manifests itself in that aftereffects of ambiguous sounds will be bigger than that of non-ambiguous sounds.

In the 2 (Age-group five- or eight-years old) x 2 (Adapter Sound Ambiguous or Non-ambiguous) x 2 (Identity of the Adapter /aba/ or /ada/) x 3 (Test Token A?-1, A?, A?+1) overall ANOVA on the proportion of /aba/ responses, there was a main effect of the identity of the adapting stimulus, $F(1,36) = 16.50$, $p < .001$, because there were more /aba/ responses following exposure to /aba/-adapters than /ada/-adapters, and a main effect of test token, $F(2,72) = 49.37$, $p < .001$, because there were more /aba/ responses for the /b/-like test token A?-1 than for the /d/-like test token A?+1. There was, furthermore, an interaction between test token and age-group, $F(2,72) = 11.75$, $p < .001$. Inspection of the figures shows that the eight-year-olds had steeper identification functions than the five-year-olds, indicating that the eight-year-olds responded in a more 'categorical' way. Most importantly, there was a significant interaction between age group, ambiguity of the adapter sound, and identity of the adapter stimulus, $F(1,36) = 6.48$, $p < .015$. For ease of interpretation, we computed aftereffects (Table 1) by subtracting the proportion of /aba/ responses following exposure to A?Vd from A?Vb (for ambiguous

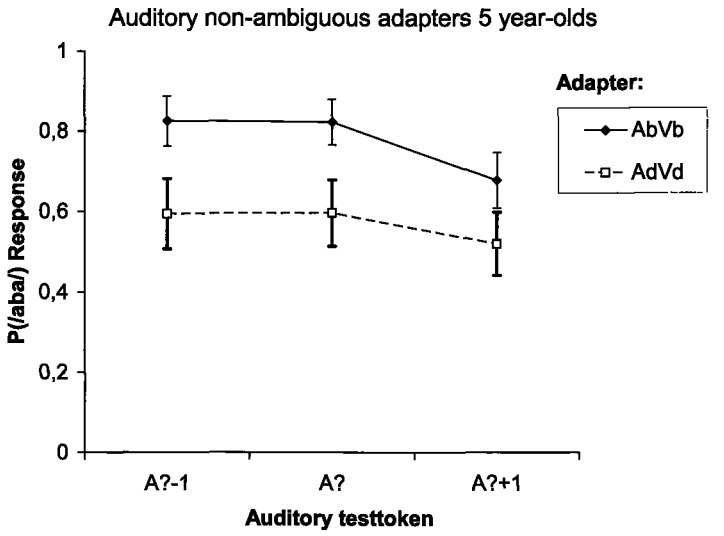
sounds) and AdVd from AbVb (for non-ambiguous sound). Separate paired t-tests confirmed that the aftereffects of the eight-year-olds for ambiguous sounds were larger than those of non-ambiguous sounds, .23 versus .08, respectively, $t(1,19) = 2.64$, $p = .016$, while for the five-year-olds, there was no such difference, .13 versus .21, respectively, $t(1, 17) = 1.12$, $p = .28$. The eight-years olds thus showed the predicted recalibration effect in that their aftereffect of ambiguous sounds was bigger than that of non-ambiguous sounds (a .15 difference in the predicted direction), while the five years-olds showed no sign of recalibration (a non-significant .08% in the opposite direction).

Figure 1



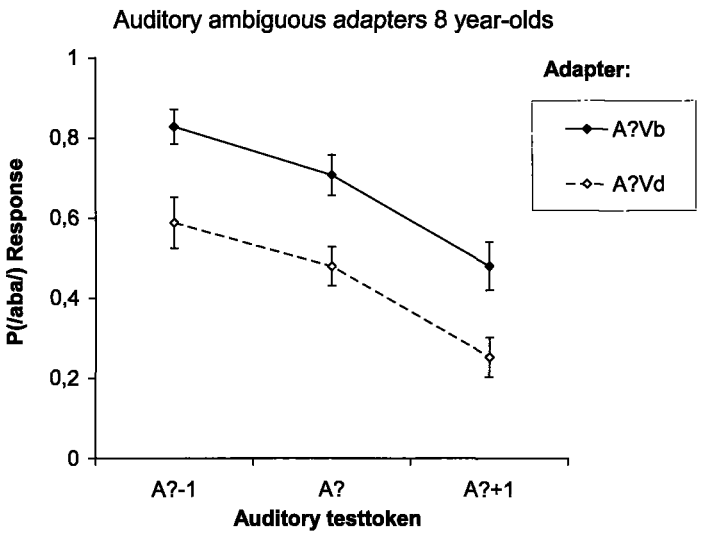
Results of the auditory posttest for the 5-year-olds. The graph shows the proportion of /aba/ responses after exposure to an adapter consisting of an ambiguous auditory token (A?) combined with either visual /aba/ (A?Vb adapter) or visual /ada/ (A?Vd adapter). On each posttest, the presented auditory token was either the ambiguous token (A?), or one of its two neighbors on the auditory continuum (A?-1, or A?+1).

Figure 2



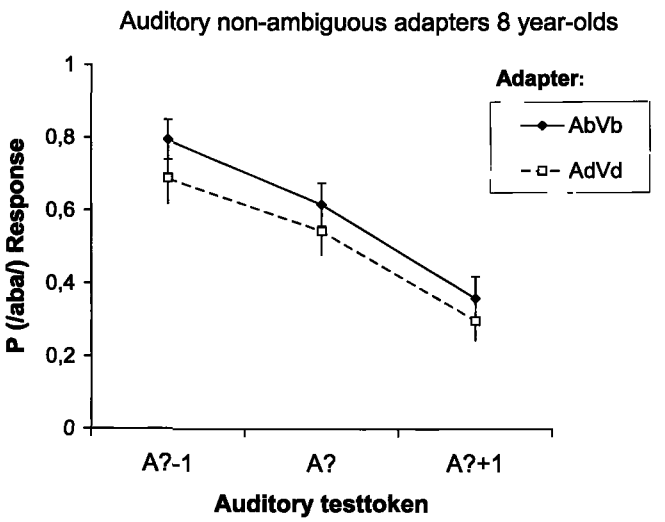
Results of the auditory posttest for the 5-year-olds. The graph shows the proportion of /aba/ responses after exposure to an adapter consisting of a non-ambiguous auditory token /aba/ or /ada/ combined with the congruent visual stimulus /aba/ (AbVb adapter) or /ada/ (AdVd adapter). On each posttest, the presented auditory token was either the ambiguous token (A?), or one of its two neighbors on the auditory continuum (A?-1, or A?+1).

Figure 3



Results of the auditory posttest for the 8-year-olds. The graph shows the proportion of /aba/ responses after exposure to an adapter consisting of an ambiguous auditory token (A?) combined with either visual /aba/ (A?Vb adapter) or visual /ada/ (A?Vd adapter). On each posttest, the presented auditory token was either the ambiguous token (A?), or one of its two neighbors on the auditory continuum (A?-1, or A?+1).

Figure 4



Results of the auditory posttest for the 8-year-olds. The graph shows the proportion of /aba/ responses after exposure to an adapter consisting of a non-ambiguous auditory token /aba/ or /ada/ combined with the congruent visual stimulus /aba/ (AbVb adapter) or /ada/ (AdVd adapter). On each posttest, the presented auditory token was either the ambiguous token (A?), or one of its two neighbors on the auditory continuum (A?-1, or A?+1).

Discussion

Five-year-olds and eight-year-olds were exposed to the video of a face saying /aba/ or /ada/ accompanied by auditory non-ambiguous (AbVb and AdVd) and auditory ambiguous (A?Vb and A?Vd) speech sounds. Following exposure to these audiovisual stimuli, aftereffects were measured in auditory-only speech identification trials. For the eight-year-olds, aftereffects were larger following exposure to the ambiguous sounds (i.e., more /aba/ responses following exposure to A?Vb rather than A?Vd) rather than the non-ambiguous sounds (AbVb and AdVd), while there was no difference between ambiguous and non-ambiguous sounds for the five-year-olds.

These results indicate that the eight-years-olds, but not the five-year-olds, adjusted their phoneme boundaries due to exposure to incongruent audiovisual speech tokens. This is consistent with a recalibration interpretation in which the phonetic conflict in the audiovisual stimuli A?Vb or A?Vd causes a shift in the perception of the ambiguous auditory speech tokens. This shift might only occur in the older group, but not the younger

one because lipreading is not yet very effective at the age of five. Previous studies have indeed shown that adults who are more susceptible to the McGurk-illusion will also tend to show larger recalibration effects (van Linden & Vroomen, in press). Given that the impact of visual information on speech perception increases with age (Hockley & Polka, 1994; Massaro, 1984; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2004), one might thus expect smaller recalibration effects in the younger age group. Possibly this reduced effects in the younger children can be attributed to developmental differences in visual information sensitivity rather than to differences in integration processes (Massaro, 1984).

In comparison with previous results obtained with adults (Bertelson et al., 2003), we also expected that selective speech adaptation would manifest itself as a contrastive effect following exposure to non-ambiguous speech sounds (i.e., less /aba/ responses following exposure to AbVb rather than AdVd). However, this shift was not observed in the present study. Here, we can offer two possibilities regarding the lack of a contrastive shift with non-ambiguous speech sounds. Firstly, as already mentioned in the introduction, we expected selective speech adaptation to be relatively small if compared with recalibration, because the amount of exposure trials given to children was optimal for observing recalibration, but on the small side for selective speech adaptation. Secondly, it also seems likely that children adopted a response strategy to repeat the phoneme of the foregoing exposure phase. Children thus responded more often /aba/ following exposure to 'aba'-adapters (AbVb and A?Vb) than following exposure to 'ada'-adapters (AdVd and A?Vd). This response strategy will mask and eventually overrule the manifestation of selective speech adaptation. The use of two types of adapters (non-ambiguous and ambiguous speech sounds), though, allowed us to subtract out the contribution of such a deliberate response strategy. By so doing, we clearly observed that for the eight-year-olds aftereffects of ambiguous and non-ambiguous sounds differed in the expected direction, thus indicating that there was recalibration, while there was no sign of this for the five-year-olds.

Our results may also shed new light on previous studies showing that adults have larger selective speech adaptation effects than children (Sussman & Carney, 1989). This has been attributed to the fact that adults are more categorical in perceiving the test stimuli, a phenomenon we also observed with the older children (see also Hazan & Barrett, 2000; Morrongiello, Robson, Best, & Clifton, 1984; Nittrouer & Studdert-Kennedy, 1987). The present results, though, make it clear that in order to measure selective speech adaptation, a baseline is needed that takes into account the various response strategies that participants might use. One such strategy might be simply to repeat the exposure stimulus during the test. For future research on selective speech adaptation, it seems important to include a baseline that takes such strategic effects into account, especially

when comparing different groups because they might differ on these strategies.

In conclusion, here we demonstrated that eight years-olds, but not five years-olds learned to categorize ambiguous speech sounds by the use of concurrently presented lipread information. The five years-old may not yet be proficient in adjusting phonetic boundaries because they are not adept decoders of lipread information. Further testing is needed to explore how phonetic skills in lipreading relate to phonetic learning, and how these skills might develop.

Chapter 7

Recalibration of phonetic categories by lipread speech versus lexical information.

van Linden, S., & Vroomen, J. (in press). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, x,* xxx-xxx.

Introduction

Frequently we encounter speakers with unfamiliar accents who are difficult to understand. But, in natural speech there are other information sources that can help listeners by telling how to interpret speech sounds that initially might be ambiguous. Two potentially important information sources are visual information from the articulators of the face – here referred to as lipread speech - and lexical knowledge. As an example, imagine an unknown 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 might have noticed that the lips were closed during pronunciation of the ambiguous sound, which is typical for /b/ but not /d/. Lexical knowledge also informs 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). What is less known, though, is that next time listeners hear the same sound, they may have learned from the past and now perceive the initially ambiguous "b/d" sound as /b/ right away (Bertelson et al., 2003; Eisner & McQueen, 2006; Kraljic & Samuel, 2005, 2006a; Norris et al., 2003). The occurrence of such an aftereffect is taken as an indication that listeners have adjusted the phonetic categories of their language so as to adapt to the new situation. Here we address whether the phonetic adjustment - or recalibration – differs when it is evoked by lipread information versus lexical knowledge. The following sections provide brief reviews of the evidence for lipread and lexical recalibration, and a discussion of why the two might differ.

Recalibration of phonetic boundaries by lipread speech has first been demonstrated by Bertelson, Vroomen and de Gelder (2003). They presented participants an ambiguous sound intermediate between /aba/ and /ada/ dubbed onto a face articulating either /aba/ or /ada/. After exposure to the auditory ambiguous speech sound combined with a face articulating /aba/, participants reported more /aba/ responses on subsequent auditory categorization trials if compared to exposure to a face articulating /ada/. The occurrence of such an aftereffect demonstrates the basic recalibration phenomenon. That is, listeners learned to categorize the ambiguous sound in accord with the lipread information it was previously combined with. A control experiment also demonstrated that when the sound was a non-ambiguous /aba/ sound dubbed onto the congruent articulatory gestures, there were less /aba/ responses on subsequent categorization trials of the auditory ambiguous stimulus if compared with exposure to non-ambiguous /ada/, thus revealing selective speech adaptation (Eimas & Corbit, 1973). Selective speech adaptation reveals fatigue of some of the relevant processes (Eimas & Corbit; Samuel, 1986), and strongly depends on repeated exposure to non-ambiguous

speech sounds. Importantly, subsequent tests showed that participants were unable to distinguish the ambiguous from non-ambiguous exposure stimuli when asked to do so in a discrimination task, thus excluding the possibility that the results were caused by deliberate response strategies (Bertelson et al., 2003).

Recalibration driven by lexical knowledge has first been demonstrated by Norris, McQueen and Cutler (2003). 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 a /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 pseudowords, 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 /S/ was ambiguous (halfway between /s/ and /S/). 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 listeners learned to interpret an ambiguous sound as /f/ or /s/. Results showed that perceptual adjustment measured after 12 h was as robust as measured immediately after learning. Equivalent effects were found when listeners heard speech from other talkers in the 12 h interval or when they could sleep.

At first sight, it may seem that phonetic recalibration driven by lipread and lexical information is much alike. Both potentially rely on the same mechanism in the sense that the phonetic boundary between two speech categories is adjusted in accordance with disambiguating information that tells what the sound should be. Whether the information stems from lipread speech or lexical knowledge might be immaterial from this point of view. There are, however, potentially important differences between the two information sources that justify further exploration. One is concerned with the fact that lipreading, by its nature, is very different from lexical knowledge; the other is that studies on lipread recalibration have found more transient effects than those on lexical recalibration. Both issues are dealt with in the following sections.

In the literature on speech perception, there are two sets of theories that explain the roles of lipreading (Green, 1998; Massaro, 1987; Summerfield, 1987) and lexical knowledge (McClelland & Elman, 1986; Norris et al., 2000) in speech perception. There is, however, very little cross-talk between the two, and both have largely been developed independently of one another. As concerns lipreading, numerous studies have shown that

seeing a person speak has a profound effect on speech perception. One of the most striking demonstrations is the McGurk effect where listeners report to hear /da/, when in fact they are presented to auditory clear /ba/ combined with a face articulating /ga/ (McGurk & MacDonald, 1976). Although there is some debate about whether integration of the auditory and visual signal occurred early or late (Schwartz et al., 1998; Vroomen, 1992), it seems clear that listeners integrate the two information sources at or before phonetic classification.

Whereas the use of so-called 'bottom-up' lipread information in speech perception is undisputed, the status of 'top-down' lexical effects remains much more debated. There are numerous studies that have shown that the lexical status of an utterance (whether it is a word or not) matters to speech perception. This has been demonstrated in paradigms like phoneme categorization (Ganong, 1980), phoneme monitoring (Cutler et al., 1987b), and phoneme restoration (Samuel, 1981b). As an example, Ganong observed that an ambiguous phoneme between /d/ and /t/ preceding 'ASK' tended to be categorized as /t/ presumably because TASK but not DASK is a word (if compared to ASH, see also Pitt, 1995). It is not clear, though, whether this shift reflects a genuine perceptual phenomenon or a response bias, and some have taken the stance that lexical knowledge is actually not used in 'on-line' speech processing proper (Norris et al., 2000).

Besides this potential difference between lipread speech and lexical knowledge in the on-line processing of speech sounds, there are other distinctions that might be crucial. Developmental studies have shown that there is a close link between lipreading and speech perception from very early on in life (Kuhl & Meltzoff, 1996), while for lexical information, it seems logical that it can only start to emerge when the lexicon starts to develop. Lipread and lexical information also differ in the time at which the information becomes available in on-line speech processing. Due to anticipatory articulation, lipread information can be available even before the speech signal is heard (Munhall & Tohkura, 1998), while lexical effects are typically slower and are usually only obtained after the word is recognized (Fox, 1984; Pitt, 1995). Lipreading can also result in stronger effects on speech perception than lexical information. For example, whereas lexical effects are typically found with ambiguous or degraded speech (Ganong, 1980; David H. Warren, 1970a), lipread information can alter the perception of auditory clear speech sounds as demonstrated in the McGurk effect (McGurk & MacDonald, 1976).

The previous examples might lead one to expect that lipreading will have more profound effects on speech perception than lexical information, as only lipreading provides perceptual and anticipatory information about the ongoing speech signal with a strong impact on the perceived sound. Such a pattern was indeed observed in a study by Brancazio (2004) who compared lipread and lexical effects on phoneme categorization. He observed that the effects of lipreading were strong and present throughout in slow,

medium, and fast responses if compared to lexical effects that were smaller and only reliable in the medium and particularly slow responses. To the extent that these categorization effects relate to recalibration, one might also expect that lipread recalibration is stronger and more robust than lexical recalibration.
Surprisingly, though, this does not seem to be the case. When comparing studies on lipread and lexical recalibration, it appears that lipread aftereffects are more fragile. For example, Vroomen et al. (2004) reported rapid dissipation of lipread aftereffects with prolonged testing. They presented 50 audiovisual exposure stimuli followed by 60 post-test trials and observed that aftereffects had dissipated after only six post-test trials.

Contrary to the fast dissipation of lipread aftereffects, studies on lexical recalibration report stable effects over time. For example, after exposure to only 20 ambiguous speech sounds, Kraljic and Samuel (2005) observed no decline in the magnitude of lexical aftereffects when post-tests were presented either immediately or after 25 minutes. Lexical aftereffects were also resistant to various types of unlearning conditions in the 25-minute interval, and they only diminished when participants heard the critical phoneme spoken unambiguously by the speaker of the exposure phase. Even larger intervals were used by Eisner and McQueen (2006) who found lexical aftereffects to remain stable over a period of even 12 hours subsequent to exposure to only a short story.

For the time being, it is unclear why lipread aftereffects seem to dissipate faster than lexical aftereffects because the various studies are difficult to compare. They not only differ in the use of lipread versus lexical information, but also in the nature of the phonemes (syllable-initial stops versus word-final fricatives or stops) and various other experimental procedures. For example, studies on lexical recalibration typically presented during the exposure phase not only the ambiguous sound that presumably drives recalibration (e.g., witlo?), but also the unambiguous sound from the opposite phoneme category (e.g. radijs). At present, it is unknown whether the presence of this sound boosts aftereffects, for example by enhancing the contrast. Furthermore, the short-lived lipread aftereffects reported by Vroomen et al. (2004) are, in principle, not mutually exclusive with the long-lasting lexical aftereffects reported by others (Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Norris et al., 2003) because lipread aftereffects have hitherto not been analyzed as how they survive a long silent period.

In the present study, we tried to resolve these issues by directly comparing aftereffects evoked by lipread and lexical information using the same paradigm and the same test stimuli. Experiment 1 served as a check that our newly created lipread and lexical exposure stimuli did indeed induce a bias effect in phoneme categorization and corresponding aftereffects. Experiment 2 explored dissipation of lipread and lexical aftereffects by measuring them as a function of the serial position in the post-test. Experiment 3 explored whether the presence of a contrast phoneme in the exposure phase

enhanced aftereffects. In Experiment 4, a 3-min silent interval was introduced between the exposure phase and the post-test to check whether aftereffects dissipate if no testing intervenes, and Experiment 5 tested whether it mattered whether participants were exposed to one versus both phoneme categories.

Experiment 1

Lipread and lexical exposure stimuli were created that biased an ambiguous sound halfway between /t/ and /p/ towards either /t/ or /p/. In the critical part of the experiment, participants were exposed to these stimuli (i.e., lipread or lexical t- or p-exposure) for a short time, immediately followed by a short post-test phase in which auditory tokens near the phoneme boundary were categorized as /t/ or /p/. These exposure - post-test phases were presented several times during the experiment with exposure towards /t/ or /p/ in random order. Experiment 1 served as a check that the exposure stimuli did indeed induce a bias effect on categorization and a corresponding aftereffect. The experiment also allowed us to compare the size of lipread and lexical aftereffects. It was expected that lipread speech would evoke bigger bias effects on categorization than lexical information as it is known that lipreading can affect even auditory clear speech tokens, whereas lexical effects are typically only obtained when the auditory stimuli are ambiguous. No predictions could be made concerning the magnitude of lipread and lexical aftereffects because relevant data were not available.

Method

Participants

Twenty-nine first-year students, all native speakers of Dutch, participated in return of course credits.

Materials

An auditory ambiguous sound intermediate between /t/ and /p/, henceforth /?/, was created using the Praat speech editor (http://www.praat.org). For the effects of lipreading, the ambiguous sound /?/ was embedded in pseudowords like /wo?/ and dubbed on the video of a face that articulated either /wop/ or /wot/. For inducing lexical effects, /?/ was embedded in a context that made up a 't'-or 'p'-word. As an example, when /?/ was embedded in 'groo?', it made up a 't'-word because 'groot' (big), but not 'groop' is a word in Dutch. Similarly, when /?/ was embedded in 'knoo?', it made up a 'p'-word because 'knoop' (button), but not 'knoot' is a word in Dutch. Eight words and eight pseudowords were recorded; four ending in /ot/ and four in /op/. Words and pseudowords were matched on number of syllables (mono, bi or trisyllabic), and they contained no other

instances of /p/ or /t/. The average (logarithmic) frequency of occurrence (per million words) of the t-words was 1.066, and for p-words it was .764 (Celex Lexical Database, 1993). An overview of the exposure stimuli is provided in table 1.

A male native speaker of Dutch was recorded on digital audio and video tape (Philips DAT- recorder and Sony PCR-PC2E mini DV). The /?/ was created from another 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 10 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.

The lexical exposure stimuli were created by excising naturally produced /op/ and /ot/ portions from words and replacing it with the synthesized token /o?/. This resulted in t-words like 'groo?', and p-words like 'knoo?'. For the lipread exposure stimuli, pseudowords were used like 'woo?' (i.e., neither 'woop' nor 'woot' is a word in Dutch) dubbed onto the video of the speaker pronouncing 'woop' or 'woot'. The video showed the speaker's face up to his eyes so that the speaker's mouth, mandible and cheeks were visible. Videos were digitized at 352 x 288 pixels at twenty-five frames per second. All video-fragments had a ten frame (250 ms) fade in and fade out with the natural synchronization between audio and video left intact. The post-test trials were made by replacing the naturally produced /ot/ from the pseudoword /sot/ with tokens from the /ot/-/op/ continuum so that listeners heard a continuum that varied from /sot/ to /sop/. Participants were tested individually in a soundproof booth. The videos were presented on a 17-inch monitor connected to a computer. The video filled about one third of the screen (10 x 9.5 cm.), and was surrounded by a black background. The sound was presented via a Fostex 6301B speaker placed underneath the monitor. Loudness was 70 dBa when measured at ear level. Participants were seated at a distance of 60 cm in front of the screen.

Table 1

Overview of the Exposure Stimuli.

	Auditory	Lipread Information
Lipread exposure stimuli	foo?	foop
		foot
	woo?	woop
		woot
	kafoo?	kaffoop
		kaffoot
	dikasoo?	dikasoop
		dikasoot
	Auditory	**Lexical Information**
Lexical exposure stimuli	knoo?	knoop (knot)
	hoo?	hoop (hope)
	siroo?	siroop (syrup)
	microscoo?	microscoop (microscope)
	groo?	groot (big)
	vloo?	vloot (fleet)
	devoo?	devoot (devout)
	idioo?	idioot (idiot)

Procedure

Testing involved four phases: a calibration phase, a training phase, an exposure phase intermitted by post-test trials for testing visually or lexically induced recalibration, and a categorization phase to examine the goodness ratings of the exposure stimuli.

Calibration: The most ambiguous token of the /sot/-/sop/ continuum was, for each individual listener, determined in the calibration phase. All tokens of the /sot/-/sop/ continuum were presented ten times in random order at 1.5 s ITI. Participants pressed a 'P' or 'T' on a keyboard upon hearing /sop/ or /sot/, respectively. The obtained s-shaped identification curve was then fitted with a logistic procedure, and the item nearest to the 50% crossover point served as the participants' most ambiguous stimulus /?/ during subsequent exposure and testing.

Training: In the training phase, participants were acquainted with the post-test procedure. The three most ambiguous tokens, the boundary token /?/ and the two tokens nearest to the boundary token, /?-1/ for the more 'p'-like token and /?+1/ for the more 't'-like token, were presented for identification. Each of the three tokens was presented twenty times with 1.5 s ITI in pseudo-randomized order. Responses were given as before.

Exposure-post-test: The exposure-post-test phase consisted of two sessions: one for testing lexical recalibration and one for lipread recalibration. Session order was counterbalanced over participants. Within each session, ten exposure-post-test blocks were presented, five blocks biasing towards /p/ and five blocks biasing towards /t/. The p- or t-blocks were presented in quasi-randomized order with no more than 2 successive p- or t-blocks in a row. One exposure - post-test block consisted of eight exposure stimuli (500 ms ITI), immediately followed by six post-tests trials (1.5 sec. ITI). The eight exposure stimuli consisted of two presentations of each of the four different exposure stimuli that biased towards either /p/ or /t/ (presentation order of the exposure stimuli counterbalanced). The six identification post-tests trials consisted of two triplets of the individually determined three most ambiguous tokens of the /sot/-/sop/ continuum (/?-1/, /?/, and /?+1/). Presentation order of the test tokens was counterbalanced so that each test token occurred equally often on each of the six serial positions of the post-test.

During exposure, participants were given no phonetic task, but to ensure that they were looking at the video during lipread exposure, they had to monitor the face for the occasional appearance of a small white dot (100 ms) on the upper lip of the speaker (catch trial). During lexical exposure, participants were viewing a white fixation-cross against a black screen. On catch trials, the fixation cross changed into the small dot (100 ms). Participants pressed a special key upon detecting a catch trial. Each session contained eight catch trials.

Rating of the exposure stimuli: In the final part of the experiment, participants rated, on a seven-point Likert-scale, the /p/-/t/ quality of the /?/ as embedded in the lexical and lipread exposure stimuli. They were asked to circle 1 upon hearing a clear /t/, 7 upon hearing a clear /p/, and 4 when indecisive about the identity of the consonant. The stimuli were presented in two blocks (lipread and lexical), each block containing 5 repetitions of each of the 8 exposure stimuli.

Results

The participants' most ambiguous token ranged from token 3 to 7 on the 10 - point continuum. Participants detected 82% of the catch trials indicating that they were looking at the screen during exposure.

Bias: We first examined whether the ambiguous phoneme /?/ was perceived as intended when embedded in the lipread and lexical exposure stimuli. The bias effect was calculated by taking the difference in the ratings of /?/ when embedded in t- versus p- stimuli. The averaged ratings were 2.15 and 6.15 for lipread t- and p-stimuli, and 4.11 and 5.32 for lexical t- and p-stimuli, respectively. In a 2 (Lipread vs. lexical information) x 2 (/p/- or /t/-stimulus) ANOVA on the ratings, there was an overall difference between lipread versus lexical stimuli, $F(1, 28) = 8.17$, $p < .01$, as the lipread /?/ was rated as

more t-like than the lexical /?/. More importantly, there was a main effect of /p/- versus /t/-stimuli, $F(1, 28) = 175.37$, $p < .001$, because /?/ was rated as more t-like when embedded in t-stimuli than p-stimuli. This bias effect interacted with information type, $F(1, 28) = 59.75$, $p < .001$, indicating that the lipread stimuli induced bigger bias effects (i.e., the 4.2 difference in the ratings of /?/ when embedded in t- versus p-stimuli) than the lexical stimuli (a 1.21 difference). Separate t-tests confirmed that lipread bias, $t(1,28) = 13.61$, $p < .001$, and lexical bias, $t(1,28) = 5.09$, $p < .001$, were both significantly bigger than zero.

Aftereffects: Post-test trials were likely to be labelled in accord with the previously presented exposure stimuli (p- or t-exposure), thus showing that there was indeed lipread and lexical recalibration. To compute aftereffects, the mean percentage of "T" responses was calculated on the post-test trials for each exposure stimulus, pooling over the three different test-stimuli (/?-1/, /?/, and /?+1/). Aftereffects were then calculated by subtracting the percentage of T-responses following p–exposure stimuli from t-exposure stimuli. For lipread stimuli, the thus computed aftereffect was 20%, for lexical exposure stimuli it was 9%. An ANOVA showed that aftereffects were, in general bigger than zero, $F(1, 28) = 37.53$, $p < .001$, and that even though aftereffects induced by lipread exposure stimuli were numerically bigger, they were not significantly different from lexical aftereffects, $F(1, 28) = 2.56$, $p = .121$.

We also explored whether there was a relation between the size of the bias and the aftereffect. There was a general tendency that participants with large bias effects had also large aftereffects. For lipread stimuli, the correlation just failed to reach significance, $r(n=29) = .317$, $p = .094$, while it was significant for lexical stimuli, $r(n=29) = .437$, $p < .02$. When the size of the bias was entered as a covariate in the comparison of lipread versus lexical aftereffects, there was no sign that lipread and lexical aftereffects were different from each other ($F < 1$). The size of lipread and lexical aftereffects was thus comparable, in particular if the difference in bias was taken into account.

Discussion

Exposure to lipread and lexical stimuli resulted in aftereffects that were comparable in size. The aftereffects could be interpreted as the manifestation of recalibration because the ambiguous sound was identified in accord with previously seen (lipread) or heard (lexical) information. The information that induces the shift can thus be bottom-up lipread information or top-down lexical knowledge. As predicted, lipread information resulted in a stronger bias effect on phoneme categorization than lexical information, but at this stage, there is no reason to maintain that there is a difference in the size of the aftereffects induced by the two information sources. In the following

experiments, we explored whether lipread and lexical aftereffects would last equally long by measuring dissipation of their aftereffects.

Experiment 2

Experiment 2 explored other potential differences between aftereffects induced by lipread and lexical information. One such difference is the rate at which the two effects dissipate. Previous studies have shown that lipread aftereffects dissipate quickly, whereas lexical aftereffects seem to last much longer (Eisner & McQueen, 2006; Kraljic & Samuel, 2005, Vroomen et al., 2004). Here, we compared the two directly by measuring dissipation of lipread and lexical aftereffects over the course of prolonged testing.

Method

Thirty new first-year psychology students participated in the experiment. All were native speakers of Dutch. Stimuli and procedures were as in Experiment 1, except that the number of post-test trials was increased from 6 (in Experiment 1) to 60, thus allowing the measurement of dissipation of lipread and lexical aftereffects. The 60 post-test trials consisted of twenty triplets of the participants three most ambiguous stimuli (/?-1/, /?/, and /?+1/) presented in counterbalanced order. Testing lasted approximately 2.5 hours with regular pauses interspersed.

Results

The participants' most ambiguous token varied from token 3 to 7. On average, 96% of the catch trials were detected.

Bias: The average ratings of the /?/ phoneme when embedded in t- and p-exposure stimuli was 2.03 and 5.92 for the lipread stimuli, and 4.49 and 5.36 for lexical exposure stimuli, respectively. In a 2 (Lipread vs. lexical information) x 2 (/t/- or /p/-exposure stimulus) ANOVA, the effect of information type, $F(1,29) = 18.07$, $p < .001$, the effect of exposure stimulus, $F(1,29) = 225.51$, $p < .001$, and the interaction, $F(1,29) = 81.02$, $p < .001$, were significant. As in Experiment 1, both lipread and lexical stimuli induced bias effects with lipread stimuli being more potent than lexical stimuli (3.89 vs. .88). Separate t-tests confirmed that both effects were bigger than zero: $t(1, 29) = 17, 45$, $p < .001$ for lipread stimuli and $t(1,29) = 3.67$, $p < .005$ for lexical stimuli.

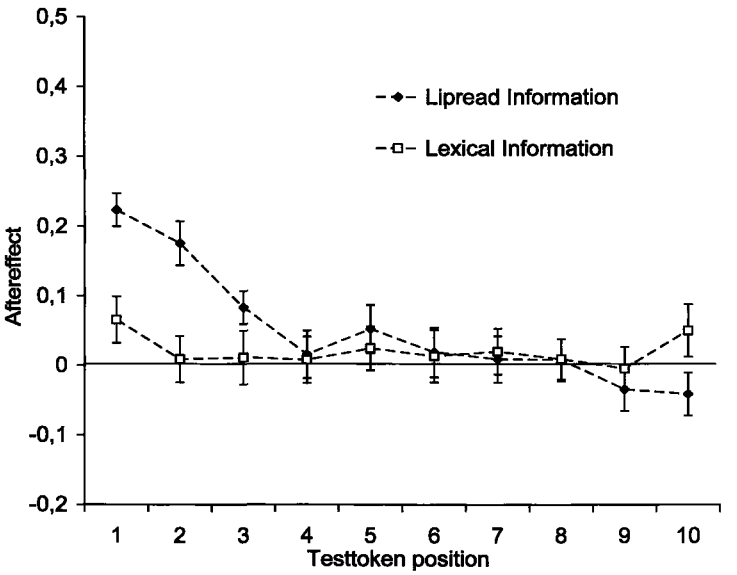
Aftereffects: In order to measure dissipation of lipread and lexical aftereffects, responses on the 60 post-test trials were first binned into 10 serial positions, with each position representing the average number of "T" responses on 6 consecutive post-test trials. Aftereffects were then calculated by subtracting the proportion of "T" responses following p-exposure from t-exposure for each of the ten serial positions. Figure 1 shows

the thus computed aftereffects, with positive numbers reflecting more responses consistent with the previous exposure stimuli (i.e., recalibration). As is clear from this figure, both lipread and lexical exposure stimuli induced positive aftereffects, but only on the first serial positions of the test. Lipread aftereffects started out stronger and lasted somewhat longer (up to the third serial position, or until post-test trials 13-18) than lexical aftereffects (lasting up to the first serial position only, or post-test trials 1-6), but both effects dissipated quickly. There was thus no sign that lexical recalibration would last longer than lipread recalibration.

A 2 (Lipread vs. lexical Information) x 10 (Test token position) ANOVA on the aftereffects showed that, on average, aftereffects were bigger than zero, $F(1, 29) = 7.23$, $p < .015$, thus indicating that there were more "T" responses following t-exposure than p-exposure. There was no overall difference in the size of lipread and lexical aftereffects, $F(1, 29) = 1.24$, $p < .274$. The main effect of test token position was significant, indicating that aftereffects became smaller when more testtrials were presented, $F(9, 261) = 5.60$, $p < .001$. The interaction between information type and test token position was significant, $F(9, 261) = 4.61$, $p < .001$. Separate t-tests showed that lipread aftereffects were significantly bigger than zero up to serial position 3 (i.e. until test trials 13-18; all p's at least < .05), whereas lexical aftereffects were significant only at serial position 1 (i.e. until test trial 1-6). Paired t-tests also showed that lipread aftereffects were bigger than lexical aftereffects on serial position 1, $t(1,29) = 3.88$, $p = .001$, and serial position 2, $t(1,29) = 3.57$, $p = .001$.

The correlation between the amount of bias in the categorization responses and the aftereffects (on the first serial position only) was not significant for lipread stimuli, $r(n=30) =. 078$, $p =.68$, and marginally so for lexical stimuli, $r(n=30) =.339$ $p = .067$. The difference between lipread and lexical aftereffects on the first and second serial position was not significant anymore, $F(1,28) = 2.10$, $p = .158$, when the difference in bias (3.88 vs. .87) was entered as a covariate in a 2 (Lipread vs. lexical Information) x 2 (Test token position) ANCOVA, indicating that the magnitude of the lipread and lexical aftereffects was comparable in size if the difference in bias was taken into account.

Figure 1



Aftereffects observed in Experiment 2 as a function of the serial position in the post-test. Following exposure to an auditory ambiguous speech token combined with lipread or lexical information, the proportion of responses consistent with the lipread or lexical information increased (= recalibration) on the first test token positions of subsequently delivered post-test trials.

Discussion

Lipread and lexical exposure stimuli evoked aftereffects that dissipated quickly with prolonged testing. Contrary to expectations based on previous studies (Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Vroomen et al., 2004), there was no sign that lexically induced aftereffects lasted longer than lipread aftereffects. If anything, lipread aftereffects tended to last somewhat longer, a result that was well accounted for by the fact that lipread stimuli also exerted a stronger bias effect than lexical stimuli. These results suggest that there is no fundamental difference in the duration of lipread and lexical aftereffects. However, they also leave unexplained why others reported lexical aftereffects to last much longer than the ones observed here (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). In the following experiment, we therefore explored whether the presence of contrast stimuli in the exposure phase, as frequently used by others, enhance aftereffects.

Experiment 3

A potentially relevant difference between studies exploring lipread and lexical aftereffects concerns the use of contrast stimuli. Studies reporting long-lasting lexical aftereffects presented during the exposure phase not only words with ambiguous sounds, but also filler words with non-ambiguous sounds taken from the opposite side of the phoneme continuum (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). For example, in the exposure phase of Norris et al. (2003) in which an ambiguous s/f sound was biased towards /s/, there were not only exposure stimuli like "radij?" which supposedly drive recalibration (i.e., the lexical information teaches that the 's/f' sound is /s/ rather than /f/), but also contrast stimuli containing the non-ambiguous sound /f/ (e.g., "witlof"). The presence of these contrast phonemes might, possibly, cause selective speech adaptation. For example, the /f/ in "witlof" may cause a 'fatigue' of /f/-detectors such that an ambiguous 's/f' sound is heard as 's' as well. Both "radij?" and "witlof" may therefore cause a shift in the phoneme boundary such that the 's/f' sound is perceived as /s/: "radij?" via recalibration and "witlof" via selective speech adaptation. For these reasons, Norris et al. (2003) included a control condition in which it was actually checked - and discarded – that selective speech adaptation was at stake. However, since there is no inherent reason why contrast stimuli should be present in the exposure phase, it might be more appropriate to exclude them from the exposure phase right away, rather than controlling for them. Moreover, the recalibration stimuli might interact with the contrast stimuli because the two phonemes together create a contrast that can trigger criterion-setting operations resulting in long-range aftereffects. Possibly then, previous studies using contrast stimuli could have overestimated the contribution of lexical recalibration. In Experiment 3, we tested this possibility by repeating Experiment 2, but now including contrast stimuli in the exposure phase. If contrast stimuli do indeed enhance aftereffects, then lipread and lexical aftereffects might become bigger or more stable in time.

Method

Twenty-four new first-year psychology students participated. Stimuli and procedures were as in Experiment 2, except that non-ambiguous contrast stimuli were included in the exposure phase. A single exposure block contained sixteen stimuli: eight stimuli with the ambiguous phoneme that biased /?/ towards either /t/ or /p/, and eight contrast stimuli that contained clear tokens of the non-ambiguous contrast phoneme. Participants might thus hear the lexical exposure stimulus 'knoo?' (biasing /?/ towards /p/) and the contrast stimulus 'groot' (which contains the non-ambiguous sound /t/) in a single exposure block. The order of the exposure stimuli was quasi-randomized with no more than three exposure stimuli or contrast stimuli in a row.

Results and discussion

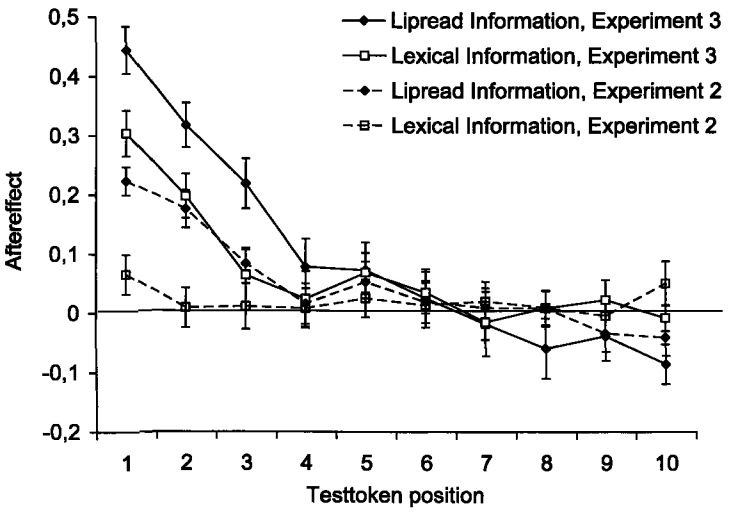
The participants' most ambiguous token varied from token 3 to 8. Participants detected on average 94% of the catch trials indicating they were attending the screen. Bias. Both lipread and lexical exposure stimuli induced a bias effect. The average rating of lipread /t/ and /p/ stimuli was 1.96 and 5.86 points, and for lexical stimuli the ratings were 3.69 points and 4.93 points, respectively. In a 2 (Lipread vs. lexical stimulus) x 2 (p- or t-stimulus) ANOVA, there was no overall difference between lipread and lexical stimuli, $F(1,23) = 3.00$, $p = .097$. The effect of t- versus p-stimulus was significant, $F(1,23) = 182.46$, $p < .001$, as was the interaction with information type, $F(1,23) = 32.41$, $p < .001$. As before, /?/ was rated as more t-like when embedded in t-stimuli than p-stimuli, and lipread information exerted a stronger bias than lexical information. Separate t-test showed that the 3.89 bias effect of lipread stimuli was significant, $t(1,23) = 12.93$, $p < .001$, as was the 1.24 effect of lexical stimuli, $t(1,23) = 4.13$, $p < .001$.

Aftereffects: Figure 2 displays aftereffects induced by lipread and lexical stimuli in Experiment 3, together with those of Experiment 2. As is clear from the figure, contrast stimuli indeed enhanced the magnitude of the aftereffects, but aftereffects still dissipated quickly. A 2 (Lipread vs. lexical Information) x 10 (Test token position) ANOVA on the aftereffects of Experiment 3 showed that aftereffects were significantly bigger than zero, $F(1, 23) = 27.99$, $p < .001$. There was no overall difference between lipread and lexical aftereffects ($F < 1$), and the effect of serial position of the test token was significant, $F(9, 207) = 34.15$, $p < .001$, as aftereffects became smaller when more test tokens were presented. The interaction between information type and test token position was also significant, $F(9, 207) = 3.52$, $p < .001$. Separate t-test showed that lipread aftereffects were bigger than zero up to test token position 3 (post-test trials 1-18) , while lexical aftereffects were significant up to test token position 2 (post-test trials 1-12; all p's < .001). There was no relation between the magnitude of the bias effect and the aftereffect on the first test token position for lipread, $r(n=24) = 016$, $p = .941$, or lexical stimuli $r(n=24) = .247$ $p = .245$.

To analyze whether the contrast stimuli boosted aftereffects, we compared aftereffects of Experiments 2 and 3 in a 2 (Experiment 2 vs. 3) x 2 (Lipread vs. lexical Information) x 10 (Test token position) ANOVA. The aftereffects of Experiment 3 were, in general, bigger than those of Experiment 2, $F(1, 52) = 5.25$, $p < .03$, demonstrating that contrast stimuli indeed enhanced the magnitude of the aftereffects. The effect of test token position was significant, $F(9,468) = 35.18$, $p < .001$, as aftereffects in both experiments dissipated with prolonged post-testing. The interaction between test token position and experiment, $F(9,468) = 8.54$, $p < .001$, was also significant indicating that aftereffects of Experiment 3 were bigger than those of Experiment 2 on the first test token positions only. Separate t-tests confirmed that lipread aftereffects were bigger in

Experiment 3 than Experiment 2 on test token positions 1-3, while lexical aftereffects were bigger on test token positions 1 and 2 (all p's < .01). Contrast stimuli with auditory non-ambiguous sounds thus enhanced the magnitude of lipread and lexical aftereffects, but only at the beginning of the test. Importantly, there was no sign that aftereffects would also become more stable when contrast stimuli were included. This issue about the stability of the aftereffects was further explored in Experiment 4.

Figure 2



Experiments 2 and 3: aftereffects as a function of the serial position in the post-test. Following exposure to an auditory ambiguous speech token combined with lipread or lexical information, the proportion of responses consistent with the lipread or lexical information increased (= recalibration) on the first test token positions of subsequently delivered post-test trials. The presence of a contrast phoneme during the exposure phase (Exp. 3) increased this effect, but aftereffects did not become more stable in time (if compared with Exp 2).

Experiment 4

It may seem that the short-lived aftereffects reported in Experiments 2 and 3 are in contradiction with long-lasting lexical aftereffects reported by others (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). In studies reporting lexical aftereffects, there is typically an interval between the exposure phase and the beginning of the post-test that may vary

from approximately 1 min (Norris, McQueen; Samuel) to 25 min (Kraljic and Samuel) up to 12 hours (Eisner and McQueen). This raises the question why aftereffects in the present study dissipate within the first minute of testing, while others observe aftereffects after so much longer times? Presumably, aftereffects dissipate because the phoneme boundary is re-adjusted back to normal. This re-adjustment, though, may occur either because time simply passes, or because listeners change their criterion while being tested. For example, one possibility is that the response criterion is adjusted in due course of testing such that the two response alternatives are chosen about equally often. There might then be no real difference between the short-lived aftereffects reported here and the stable aftereffects reported by others, because both can remain stable until the post-test phase starts.

To explore whether aftereffects indeed remain stable as long as no new test tokens are encountered, we introduced a 3-min silent interval between the end of the exposure phase and the beginning of the post-test phase. If the mere passing of time causes dissipation, one would expect aftereffects as observed in Experiments 2 and 3 to have dissipated completely following a 3-min interval. Alternatively, a silent interval might not harm aftereffects if dissipation occurs as a consequence of criterion shifts during the post-test.

Method

Twenty-nine new students participated in Experiment 4. Stimuli and procedures were as in Experiment 2, except that a 3-min silent interval was introduced between the end of the exposure phase and the beginning of the post-test trials. Participants tried to solve a Rubik's cube during this interval. A short tone, 10 s before the first post-test trial, warned participants for the upcoming test phase.

Results and Discussion

The participants' most ambiguous token varied from token 3 to 6. Participants detected on average 89% of the catch trials.

Bias: Both lipread and lexical stimuli induced bias effects. The average rating of /?/ was 2.77 and 6.09 when embedded in lipread t- and p-stimuli, and 4.82 and 5.56 for lexical t- and p-stimuli, respectively. A 2 (Lipread vs. lexical information) x 2 (p- or t-stimulus) ANOVA showed that /?/ was rated as more t-like with lipread information than with lexical information, $F(1,28) = 9.63$, $p < .005$. The effect of t- versus p-stimulus was significant, $F(1,28) = 98.26$, $p < .001$, as was the interaction $F(1,28) = 48.90$, $p < .001$, indicating that lipread stimuli induced bigger bias effects than lexical stimuli. Separate t-test confirmed that the 3.32 bias effect of lipread stimuli was significant, $t(1,28) = 12.11$, $p < .001$, as was the .74 bias effects of lexical stimuli, $t(1,28) = 2.67$, $p < .015$.

Aftereffects: Aftereffects were calculated as in Experiment 2 by subtracting the proportion of "T" responses following p-exposure from t-exposure for each of the ten serial positions. Aftereffects are shown in figure 3, together with those of Experiment 2. Importantly, aftereffects survived the 3-min silent interval, indicating that the mere passing of time did not make aftereffects disappear.

A 2 (Lipread vs. lexical information) x 10 (Test token position) ANOVA on the aftereffects of Experiment 4 confirmed that aftereffects were significantly above zero, $F(1, 28) = 4.98$, $p < .05$. There was no overall difference between lipread and lexical exposure stimuli, $F < 1$, and the effect of test token position was again significant, $F(9, 252) = 2.76$, $p < .005$, as aftereffects decreased with prolonged testing. The interaction between information type and test token position was not significant, $F(9, 252) = 1.18$, $p = .306$. Despite that the interaction was not significant; we conducted separate t-tests so that a comparison could be made with the analysis of Experiment 2. For the lipread stimuli, aftereffects were bigger than zero on serial positions 1-4 (post-test trials 1-24), while for the lexical stimuli, aftereffects were bigger than zero only on serial position 1 (post-test trials 1-6; all p's < .05).

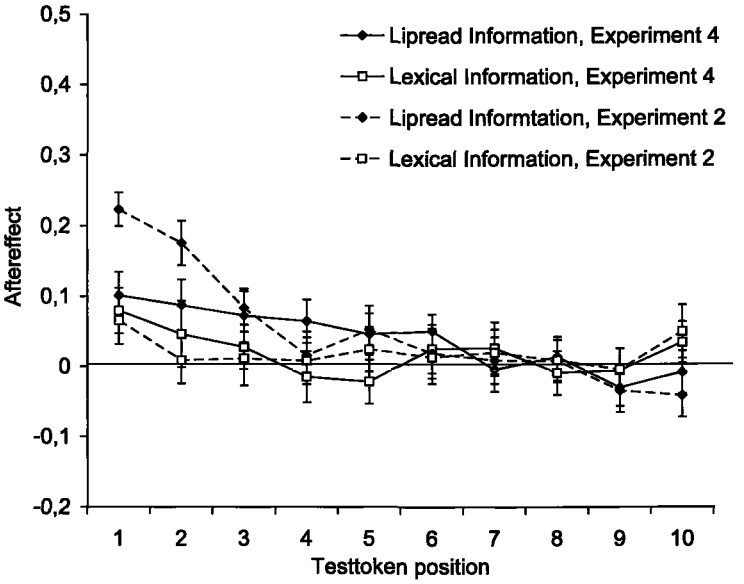
To check whether the 3-min silent interval affected aftereffects, we also compared the results of Experiment 4 with those Experiment 2 in a 2 (Experiment 2 vs. 4) x 2 (Lipread vs. lexical Information) x 10 (Test token position) ANOVA with Experiment as between-subject factor. Importantly, the main effect of experiment, and its interaction with information type and serial position were all non-significant (all F's < 1). The effect of test token position was again significant, $F(9, 513) = 7.85$, $p < .001$, as aftereffects decreased when more test tokens were presented. Inspection of figure 3 shows that there was also a tendency that lipread aftereffects following the 3-min interval were smaller on the first serial position if compared to the no-interval condition, but the second-order interaction between experiment, information type and serial position was not significant, $F(9, 513) = 1.72$, $p = .081$. There was no relation between the size of the bias effect and the aftereffect on the first serial position (post-test trials 1-6) for lipread, $r(n=29) = -.001$, $p = .997$, and lexical stimuli , $r(n=29)=.21$, $p =.275$.

However, when the same correlations were computed across the four experiments, it appeared that for the lexical stimuli, participants with big bias effects also displayed big aftereffects, $r(n=112) = .262$, $p < .01$, but there was no such relation for lipread stimuli, $r(n=112) =.068$, $p = .48$.

The results of Experiment 4 thus essentially showed that a 3-min silent interval between the exposure phase and the post-test did not make aftereffects disappear. Lipread and lexical aftereffects remained stable until the post-test phase started, and only then they disappeared quickly. In the final experiment, we further explored reasons for this dissipation. Here we addressed whether aftereffects disappear because participants in

previous experiments were exposed to both p- and t-biasing contexts, rather than just a single context.

Figure 3



Experiments 2 and 4. Aftereffects as a function of the serial position in the post-test. Following exposure to an auditory ambiguous speech token combined with lipread or lexical information, the proportion of responses consistent with the lipread or lexical information increased (= recalibration) on the first test token positions of subsequently delivered post-test trials. A 3-min pause between the exposure phase and the post-tests (Exp 4) did not affect the magnitude or the rate of dissipation of the aftereffects (if compared with Exp 2).

Experiment 5

Aftereffects as measured in Experiments 1-4 relied on the effect of the most recently encountered exposure stimuli. These exposure stimuli varied from block to block, so that participants were biased towards both phoneme categories (i.e., /p/ and /t/). This procedure allowed us to measure aftereffects in a within-subjects design. Studies on lexical aftereffects, though, have used a between-subject design and exposed participants to only a single context (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). Possibly, this difference affects the robustness of the phenomenon because in the latter cse, phoneme boundaries are not continuously adjusted to one side of the continuum or the other (but

see Vroomen et al., 2004). To check whether this difference in procedure indeed matters, participants of Experiment 5 were exposed to only a single context (i.e., lipread or lexical /p/- or /t/-exposure)

Method

Sixty new students were randomly assigned to one of four groups (15 participants per group). Each group was exposed to only one out of four possible combinations of lipread or lexical p- or t-exposure stimuli. Participants were presented five exposure-post-test blocks.

Results

The participants' most ambiguous token varied from token 3 to 7. On average, 98% of the catch trials were detected.

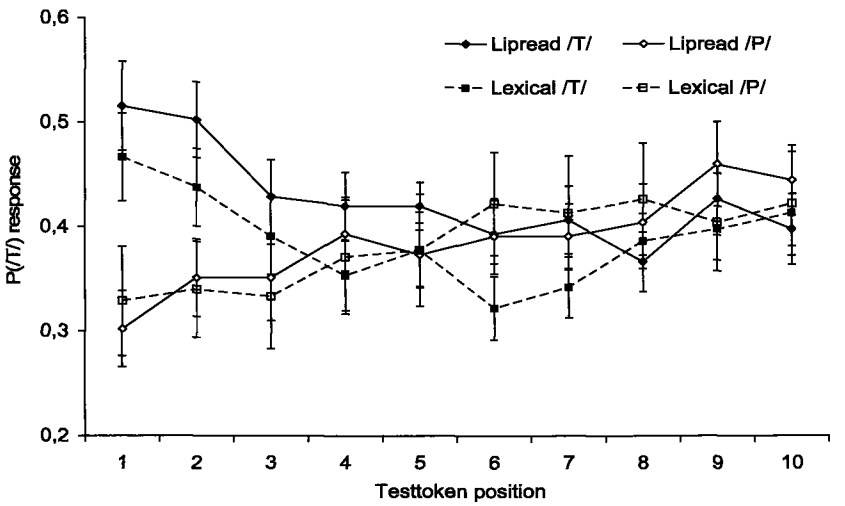
Unlike the previous experiments, aftereffects could not be computed on an individual basis because the lipread and lexical t- versus p-exposure conditions were between-subjects variables. Figure 4 therefore shows the average proportion of T-responses on the post-tests as a function of the serial position for each of the four groups separately. As is clear from this figure, in the initial phase of the post-test (up to serial position 2, or post-test trials 1-12), there were more T-responses for the /t/-exposed groups (lipread and lexical) than for the /p/-exposed groups, but this difference disappeared with prolonged testing.

This generalization was supported by a 2 (lipread vs. lexical information) x 2 (p- vs. t-exposure) x 10 (test token position) ANOVA on the proportion of T-responses with information type and exposure phoneme as between-subjects variables. In the ANOVA, there was a significant interaction between exposure phoneme and test token position, $F(9,504) = 8.26$, $p < .001$, indicating that the t-exposed groups had more T-responses than the /p/-exposed groups in the beginning of the test. There was no overall difference in number of T-responses between the lipread and lexical exposure groups, nor was any of the other interactions significant (all $F's < 1$). Separate t-tests confirmed that the lipread t-exposed group had more T responses on test token position 1 and 2 (trials 1-12) than the p-exposed group, while for the lexical t-exposed groups this was the case for serial position 1 (all $p's < .05$), and marginally so for test token position 2, $t(1, 28) = 1.66$, $p = .10$.

To examine whether aftereffects were different for participants exposed to one versus both phoneme categories (i.e., Experiments 2 vs. 5), we also computed a group averaged aftereffect for Experiment 5 by subtracting, per serial position and information type, the proportion of T-responses of the p-exposed groups from the t-exposed groups. This group-averaged aftereffect can then be compared with the individually determined aftereffects of

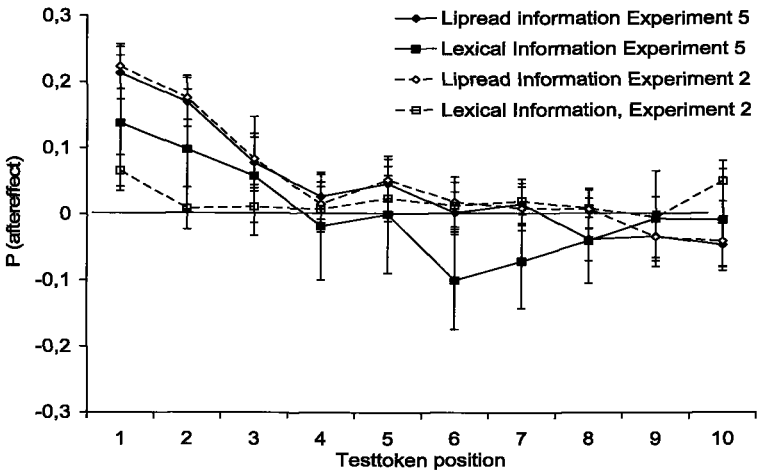
Experiment 2 (see Figure 5). As is clear from Figure 5, lipread aftereffects of Experiments 2 and 5 were very much alike: Both were significant in the beginning of the test and dissipated with prolonged testing. The lexical aftereffects of Experiment 5 were bigger than those of Experiment 2 on the first test token positions, but smaller on later positions. There was thus no sign that aftereffects in general would be bigger or last longer if participants were exposed to one instead of both phoneme categories.

Figure 4



The proportion of T-responses as a function of the serial position in the post-test as obtained in Experiment 5. There were more T-responses for the /t/-exposed groups (lipread and lexical) than for the /p/-exposed groups on the first test token positions.

Figure 5



Experiments 2 and 5: Aftereffects as a function of the serial position in the post-test. Following exposure to an auditory ambiguous speech token combined with lipread or lexical information, the proportion of responses consistent with the lipread or lexical information increased (= recalibration) on the first test token positions of subsequently delivered post-test trials. Aftereffects were not different for participants exposed to one versus both phoneme categories.

General discussion

The present study shows that when listeners hear an ambiguous phoneme, they flexibly adjust the phonetic categories of their language in accordance with disambiguating information that tells what the phoneme should be (i.e., recalibration). The disambiguating information can be lipread or lexical, and both information sources induce similar-sized aftereffects (Exp. 1). Results also show that lipread and lexical aftereffects dissipate about equally fast in due course of testing (Exp. 2). The presence of a contrast phoneme during the exposure phase enhances the size of the aftereffects (Exp. 3.), but aftereffects do not become more stable in time. Importantly, though, aftereffects do not become smaller when a 3-min silent interval intervenes between the exposure phase and the test (Exp. 4). This indicates that recalibration as such is not fragile, but that other factors possibly related to test procedure itself may explain why aftereffects dissipate quickly during testing. One such factor we further explored was whether participants were biased in consecutive exposure phases towards one or both phoneme categories. The results showed that this factor is not critical because aftereffects of participants exposed to one or both phoneme categories were comparable in size and duration (Exp. 5). Taken together,

the results show that aftereffects induced by lipread and lexical information are alike. From a functional perspective, there was thus no difference between bottom-up perceptual information or top-down stored lexical knowledge: both information sources were used in the same way to adjust the boundary between two phonetic categories.

It remains to be explained why others have observed long-lasting lexical aftereffects (Eisner & McQueen, 2006; Kraljic & Samuel, 2005), while here, we found that aftereffects dissipate fast, even if no other information is encountered that tells what ambiguous phoneme should be. One possible explanation for the fast dissipation is that listeners change their criterion while being tested. For example, it may be that listeners adjust their response criterion in due course of testing such that the two response alternatives are chosen about equally often. If so, there might be no real difference between the short-lived aftereffects reported here and the stable aftereffects reported by others, because both remain stable until the post-test phase starts. Another prediction is that if others would measure aftereffects as a function of the serial position in the post-test, they might observe that aftereffects become smaller with prolonged testing (Kraljic & Samuel, 2006). Alternatively, though, it may also be that the rate of dissipation depends on the acoustic nature of the stimuli. In the present case, syllable-final stop consonants were used that varied in place of articulation (/p/-/t/), while others used fricatives (/f-s/ and /s-S/, Eisner & McQueen, 2006; Kraljic & Samuel, 2005) or syllable-initial voiced/voiceless stop consonants (/d-t/, Kraljic & Samuel, 2006). If the stability of the phenomenon depends on the acoustic nature of the cues (e.g., a more ambiguous cue is more likely to shift than a less ambiguous cue), one may observe that aftereffects differ in this respect as well.

Another robust finding was that lipread bias on phoneme categorization was bigger than lexical bias. This result is in line with previous findings showing that lipreading exerts a stronger effect on the percept of a heard sound than lexical information does (Brancazio, 2004). What is surprising is that from the point of view where information is defined as a 'reduction in uncertainty', there is actually little reason why lipread information should have a stronger impact than lexical information, because lipread information is in fact not less constraining than lexical information. As an example, for lexical exposure stimuli like groo?, the word context specifies a single interpretation of /?/, because only 'groot' and not 'groop' is a word in Dutch. On the other hand, for lipread information one could argue that the final consonant in 'woot' is not very distinctive because /t/ is difficult to lipread (Montgomery et al., 1987). This then raises the question why lipread information, if not more constraining than lexical information, has nevertheless a stronger impact on phoneme categorization? The answer on this question probably relates to the distinction made earlier between bottom-up versus top-down information. Lexical influences will always be relatively weak and will only operate when the input signal

is weak or ambiguous. In contrast, lipread information has, from the very beginning, more weight because it is an inherent input property of the speech signal. From this point of view, then, there is a quantitative distinction between bottom-up and top-down information because lipreading outweighs lexical information in relative contribution to the immediate phonetic percept.

From a theoretical point of view, it is of interest that there was a correlation between the amount of lexical bias in phoneme categorization and the size of the lexically induced aftereffect. Listeners who displayed the biggest lexical bias also showed the biggest lexical aftereffects. This finding speaks to the hotly debated issue about whether lexical information can actually influence how people hear speech sounds. Despite the fact that numerous studies have shown that lexical information can bias phoneme categorization, an important limitation is that these studies rely on subjective reports of what listeners hear. For this reason, some have argued (Norris et al., 2000) that bias effects are not informative about speech processing proper, because they reflect decision-level influences rather than true perceptual effects.

One way to address this concern is to look for the consequences of contextual effects in a situation where listeners do not make a decision about the speech signal itself. The paradigm used here in which aftereffects were measured is an example in case. To explain the existence of (lexically induced) aftereffects, the same theorists have argued that there is a distinction between 'off-line' lexical feedback used for learning, - which is supposedly real and of benefit to speech perception -, and 'on-line' lexical feedback as embodied in interactive model of spoken word recognition – which is supposedly not real and even harmful (Norris, et al.). However, if this strict distinction between on-line and off-line feedback is valid, there is actually little reason why the two measures of these phenomena (i.e., the bias in phoneme categorization and aftereffects) should be related, given that they reflect different domains. Since our results, though, show that there is a correlation between lexical bias and aftereffects, it may be more useful to integrate the two measures into a coherent framework.

One plausible mechanism might be that listeners flexibly adjust their phoneme boundary whenever an ambiguous sound is heard in a disambiguating context. This adjustment occurs fast and instantly (Vroomen et al., 2007) , and it may for this reason show up as an immediate bias when asked to rate or identify the phoneme, while upon later testing the adjustment is still manifest to be observable as an aftereffect. Bias in phoneme categorization and aftereffects are, on this view, caused by the same mechanism and they reflect a change in the criterion of the phoneme boundary. A testable prediction that follows from this notion is that there will be no recalibration if the exposure stimuli do not evoke a bias as well.

Chapter 8

Lexical effects on auditory speech perception: An electrophysiological study.

van Linden, S., Stekelenburg, J. J., Vroomen, J., & Tuomainen, J. (in press). Lexical effects on auditory speech perception: An electrophysiological study. *Neuroscience Letters, x,* xxx-xxx.

Introduction

Identification of a speech sound is influenced by word context, especially if the sound is ambiguous or degraded. For example, an ambiguous sound that might be a /g/ or a /k/ is more likely to be identified as a /g/ if followed by 'ift' and as a /k/ if followed by 'iss'

(Ganong, 1981)Presumably, this bias effect occurs because 'gift' and 'kiss' are words in English, but not 'kift' and 'giss'. What is less known, is that next time listeners hear the same ambiguous sound, they have learned from the past and now perceive the initially ambiguous "g/k" as /g/ or /k/ right away (Kraljic & Samuel, 2005; Norris, McQueen & Cutler 2003; van Linden & Vroomen, in press). The occurrence of such a lexically-induced aftereffect is taken as an indication that listeners have adjusted, or *recalibrated*, the phonetic categories of their language so as to adapt to the new situation. Here, we explored the extent to which these phenomena are truly perceptual in nature rather than reflecting a post-lexical decision stage.

Interactive approaches argue that bias effects are due to a direct lexical influence on pre-lexical representations (McClelland, Mirman & Holt, 2006). They predict that lexical *information actually reaches down and changes the momentary activation of the sound* that is heard. Other proposals, in which speech recognition is seen as a more autonomous, bottom-up process, propose that lexical contextual information does not change the activation of the pre-lexical representation per se (Norris et al., 2003) because that will harm speech recognition proper. Lexical bias effects occur, on this view, on a post-lexical phonemic decision stage. Nevertheless, autonomous accounts leave open the possibility that pre-lexical levels are affected, but in an indirect way, via recalibration (van Linden & Vroomen, in press). The notion is that lexical information induces a shift in the boundary between two phonetic categories, and to the extent that this shift occurs at a perceptual level, one may observe that lexical information affects early processing stages. Both accounts can therefore predict that lexical information penetrates mechanisms of perception at early pre-lexical levels, and thus affecting the way a sound is heard. Here, we tested this prediction, for the first time, using recordings of human brain event-related potentials (ERPs) focusing on the mismatch negativity (MMN).

The MMN is an ERP component that signals an infrequent discriminable change in an acoustic or phonological feature of a repetitive sound (Näätänen, Gaillard & Mäntysalo, 1978). The behavioural discriminability of the stimuli is usually correlated with the *amplitude and latency of the MMN-response (Lang, Nyrke, Ek, Aaltonen, Raimo &* Näätanen, 1990). The MMN-generating process is not volitional, it does not require attentive selection of the sound (although it can be diminished under high attentional load (Sussman, Winkler, Huotilainen, Ritter & Näätänen, 2002), and it is elicited whether or not the sounds are relevant for the participant's task (Näätänen 1992; Näätänen, Paavilainen,

Tiitinen, Jiang & Alho, 1993). Furthermore, the MMN is not only sensitive to acoustic changes, but also to learned language-specific auditory deviancy (Näätänen, 2001). For example, in a cross-linguistic study of Hungarian and Finnish, Winkler et al. (Winkler, Kujala, Shtyrov, Simola, Tiitinen, Alku, Lehtokoski, Czigler, Ilmoniemi, & Näätänen, 1999) used within- and across-category phoneme contrasts that were reversed for the two languages. By means of this crossed design, they demonstrated that the MMN-generating process simultaneously operates both on the basis of auditory sensory memory and categorical phonetic stimulus representations (see also Dehaene-Lambertz, 1997; Näätänen, Lehtokoski, Lennes, Cheour, Huotilainen, Livonen, Vainio, Alku, Ilmonieme, Luuk, Allik, Sinkkonen, Alho, 1997). These results suggest that linguistic information triggers additional processes, which may prepare the auditory system for detecting language-specific auditory deviations. The pre-attentional and automatic nature of the MMN (Näätänen 1992, 1999) together with its sensitivity to phonetic contrasts and stimulus discriminability therefore makes it suitable to investigate whether lexical information can affect early pre-lexical processing stages. If it can be demonstrated that lexical information indeed changes the MMN while acoustic factors are strictly controlled for, it would naturally strengthen the idea that lexical information affects pre-lexical processes, be it direct via top-down lexical activation, or indirect via recalibration.

Here, we presented Dutch listeners a word normally ending in /t/ ('vloot', meaning 'fleet' in English) or /p/ ('hoop', meaning 'hope'), whereby the final consonant (/t/ or /p/) was replaced by an ambiguous sound halfway between /t/ and /p/ (henceforth /?/). This thus resulted in the t-word /vlo?/ and the p-word /ho?/ (note that 'vloop' and 'hoot' do not exist in Dutch). In a previous study (van Linden & Vroomen, in press), we confirmed that these words evoked a lexical bias in phoneme categorization (i.e., listeners judged /?/ in /vlo?/ as more t-like than in /ho?/) and a recalibration effect (i.e., listeners were more likely to categorize /?/ as /t/ after hearing /vlo?/ than after hearing /ho?/). The t- and p-words were presented in a typical oddball paradigm (Table 1). The standard stimulus was either the t- or the p-word containing the ambiguous sound /?/, while on infrequent deviant trials /?/ was replaced by non-ambiguous /t/. Listeners thus heard /vlot/ as deviant in the t-word condition (i.e., the word that is in congruence with the lexical information 'vloot') and /hot/ in the p-word condition (i.e., a pseudoword that is incongruent with the lexical information 'hoop'). Crucially, the acoustic change from the standard /?/ to the deviant /t/ was in the p- and t-word conditions exactly the same, as these two words only differed in their initial consonants. Interactive accounts, though, predict the MMN to be smaller in the t-word than p-word because the t-word increased activation of /t/, while the p-word increased activation of /p/. Similar predictions can be made for accounts that instantiate recalibration at an early perceptual level (van Linden & Vroomen, in press). If the shift in the phoneme boundary as evoked by the lexical

information is perceptual in nature, one expects the perceptual difference between the /?/ and /t/ to be smaller in the t-word than the p-word, because /?/ is recalibrated towards /t/, which in turn should yield a smaller MMN amplitude.

Given that the deviant in the t-word condition is a word ('vloot'), while in the p-word condition it is a pseudoword ('hoot'), one might ask on beforehand whether a smaller MMN for t-words reflects a change in the lexical status of the deviant rather than a change in the way the ambiguous sound is heard. At present there is mixed evidence about the role of the lexical status of the deviant. Pulvermüller et al. (Pulvermüller, Kujala, Shtyrov, Simola, Tiitinen, Alku, Alho, Martinkauppi, Ilmoniemi & Näätänen, 2001) argued that words engage a lexical representation in addition to the acoustic and phonetic representations activated by pseudowords, and word deviants will therefore always evoke a larger MMN than pseudoword deviants, irrespective of the lexical status of the standard. This hypothesis is thus in the opposite direction of our prediction (i.e., the t-word condition with a word as deviant will have a smaller MMN). Jacobsen et al., (Jacobsen, Horváth, Schröger, Lattner, Widmann, Winkler, 2004) though, argued that the lexical status of the deviant is irrelevant for the MMN because they found no difference between word or pseudoword deviants when acoustic and language factors were being controlled. Whichever of these two accounts is correct, here it seems safe to conclude that a smaller MMN in the t-word condition is unlikely to be caused by the fact that the deviant in this condition is a word rather than a pseudoword, because previous studies suggest that the MMN should either be bigger (Pulvermüller, et al. 2001) or not be affected (Jacobsen et al., 2004).

Method

Participants

Sixteen native speakers of Dutch (4 males, 12 females) with normal hearing and normal or corrected-to-normal vision participated in the experiment after giving written informed consent. Their age ranged from 18 to 25 years (mean age 19.5 years). The experiment was conducted in accordance with the Declaration of Helsinki.

Materials

The experiment took place in a dimly-lit, sound-attenuated, and electrically shielded room. Stimulus creation started with digital recording of /vlot/ and /hop/ by a male Dutch speaker. The final vowel and consonant of the two words were replaced by /o?/. The ambiguous sound /?/ was created with Praat (Boersma & Weenink, 2002) from another recording of /ot/ and /op/ in which the second (F2) and third (F3) formant were

changed. The steady state-value of the F2 in the vowel was 950 Hz (72 ms in duration), and the offset frequency of the transition (45 ms duration) was 928 Hz. The steady state value of F3 in the vowel was 2400 Hz, and the offset frequency of the transition was 2265 Hz. There was 40 ms of silence before the final release of the stop consonant. 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 total duration of the words were /hoo?/ = 531 ms, /hoot/ = 495 ms, /vloo?/ = 664 ms, and /vloot/ = 628 ms. Stimuli were presented from a loudspeaker located in front (90 cm) of the participant with a peak intensity of 70 dB(A).

Procedure

Testing consisted of a behavioural session and a EEG session which were run on separate days. During the behavioural session the participant's perception of the stimuli was determined. In the second session, electroencephalography was recorded to investigate whether lexical recalibration of phoneme perception is observable on the auditory evoked mismatch negativity.

During the behavioural session, we established the participant's phoneme boundary on the /soot/ - /soop/. Al 10 tokens of the continuum were presented 10 times in randomized order. Participants pressed a /p/ upon hearing /soop/ and /t/ upon hearing /soot/ on a keyboard. We also obtained a measure of lexical bias on phoneme categorization by asking participants, to rate on a 7-point Likert scale the quality of the final phoneme /?/ when embedded in the t-word /vlo?/ and p-word /ho?/.

In the MMN experiment, stimuli were presented in a typical unattended oddball paradigm (standard 82%, deviant 18%). The order of stimuli was randomized with the restriction that at least two standards preceded each deviant. The stimulus onset asynchrony was 1250 ms. During stimulus presentations, participants fixated on a small white cross on a monitor – placed directly above the loudspeaker – and detected an occasional catch trial (11% of the standard trials). Their task was to indicate by a button press when the colour of the fixation point changed. Participants were administered two blocks per word type condition, each consisting of 440 trials, which amounted to a total 720 standards (including 80 catch trials) and 160 deviants. Presentation order of the four blocks was counterbalanced across participants.

Data-analyses

The electroencephalogram (EEG) was recorded at a sample rate of 512 Hz from 43 active Ag-AgCl electrodes (BioSemi, Amsterdam, The Netherlands) mounted in an elastic cap and two mastoid electrodes. Electrodes were placed according the extended International 10-20 system. Two additional electrodes served as reference (Common Mode

Sense [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). EEG was re-referenced offline to an average of left and right mastoids and band-pass filtered (0.1–30 Hz, 24 dB/octave). The electrooculogram (EOG) measuring horizontal and vertical eye-movements was recorded using electrodes at the outer canthus of each eye as well as above and below the right eye. The raw data were segmented into epochs of 500 ms including a 100-ms prestimulus baseline. After eye movement correction [4], epochs with an amplitude change exceeding ±100 μV at any channel (except EOG) were rejected (4% of the deviant trials). ERPs of the standard and deviant non-catch trials were averaged separately for t- and p-words. The standard ERP was subtracted from the deviant ERP to obtain the MMN. To match the timing of the ERP components in the t-word condition with those in the p-word condition, ERPs were time-locked to the onset of the final phoneme of the standard and deviant word (i.e., the point where /?/ and /t/ started to deviate). Based on visual inspection of the grand average waveforms at electrode Fz, the MMN was identified as a negative deflection in a 150 - 250 ms window after the onset of the final phoneme. MMN amplitude was calculated as a 50-ms mean amplitude centred on the individual peak latency of MMN. The MMN was tested at electrodes F3, Fz, F4, FC3, FCz, FC4 with a MANOVA for repeated measures with as within subject- factors Condition (p- versus t-word), Hemisphere (left, middle, right) and Anterior-Posterior (frontal versus fronto-central). A one-tailed test for Condition was used because there was a clear prediction about the direction of the lexical effect on the MMN. Two participants were excluded from the analysis because strong alpha waves prevented reliable scoring of the MMN. Data of one participant were discarded because of hardware failure.

Table 1

Experimental design

	Standard	Deviant
t-word 'vloot'	/vlo?/	/vlot/
p-word 'hoop'	/ho?/	/hot/

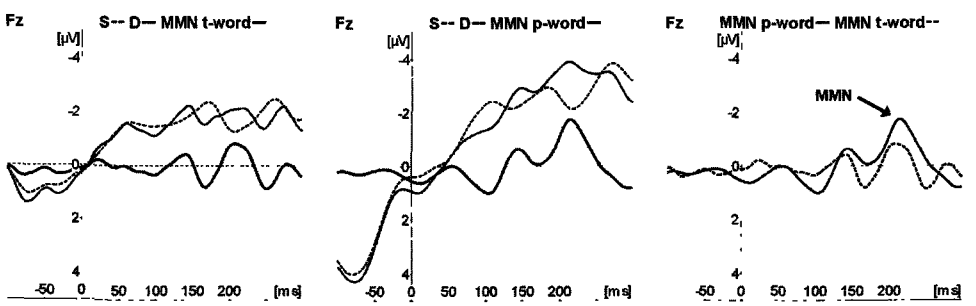
Results

Pretests confirmed that the ambiguous phoneme /?/ was perceived by each participant as halfway between /p/ and /t/ when presented in the neutral pseudoword /so?/. We also obtained a behavioral measure of lexical bias on phoneme categorization by asking participants at the end of the MMN experiment to rate on a 7-point Likert scale the

quality of the final phoneme /?/ when embedded in the t-word /vlo?/ and p-word /ho?/. As expected, /?/ in the t-word /vlo?/ was rated as more t-like than when embedded in the p-word /ho?/ (3.83 vs. 5.80 units on the 7-point Likert scale, $t(1,12) = 6.31$, $p < 0.001$).

Fig. 1 depicts the ERPs elicited by standards and deviants and the difference waves at Fz for both p- and t-conditions. MMN peaked at 215 ms at Fz with no difference between p- and t-words in their timing ($t < 1$). As predicted, the amplitude of the MMN was larger for p-words than for t-words, $F(1,12) = 3.62$, $p < 0.05$. Post-hoc analysis showed that the MMN for p-words (-1.6 μV) significantly deviated from zero, $F(1,12) = 22.74$, $p < 0.001$, whereas MMN amplitude for t-words (-0.5 μV) did not differ significantly from zero ($p = 0.18$). Testing the scalp distribution of MMN revealed no interactions between Condition and Hemisphere, Condition and Anterior-Posterior, or Condition and Hemisphere x Anterior-Posterior (all Fs < 1), indicating that the scalp distribution of the MMN did not differ between conditions.

Figure 1



Grand-averaged waveforms of the standard (S), deviant (D) and MMN at electrode Fz for the t-word condition (left panel) and p-word condition (middle panel). The right panel shows the MMN for both conditions. The y-axis marks the onset of deviation between /?/ and /t/.

Discussion

The results thus show that, with acoustic factors being controlled for, the perceptual change from /?/ to /t/ was smaller in the t-word than in the p-word. This finding is in line with accounts that attribute the lexical context effects in speech perception on a pre-lexical level rather than a post-lexical phonemic decision stage. Interactive accounts might argue that the lexical representation of the t-word 'vloot' increased the activation of /t/ via feedback connections, while the p-word 'hoop' increased

activation of /p/. A recalibration account might say that upon hearing the t-word /vlo?/, the phoneme boundary is shifted such that next time /?/ is presented, it is heard as /t/. The perceptual change from /?/ to /t/ is therefore smaller in 'vloot' than in 'hoop', eliciting on its turn an MMN of smaller amplitude.

In the present study, a visual task (i.e., detection of an occasional change in fixation) was used to draw attention away from auditory stimulation. It is unknown, though, to which extent participants ignored the auditory stimuli. Given that the MMN can be modified by attention (Sussman, 2002), future studies might manipulate the attentional load to determine whether the lexically induced MMN reflects auditory processing at a pre-attentive stage.

Future work is also needed to further elucidate the link between behavioural performance and the underlying cognitive processes and neural system that support it. For example, behaviourally, we observed that the ambiguous phoneme /?/ was rated more t-like when embedded in the t-word /vlo?/ than the p-word /ho?/. However, the size of this lexical bias effect on phoneme categorization did not correlate with the amplitude of the MMN, (r_s = -0.19, p = 0.54). There was thus no simple relation such that participants who had a large lexical bias effect also had a strong MMN. This aspect of the results will be further investigated.

Chapter 9

Summary and general discussion.

Summary and General Discussion

Perceiving speech is a complex perceptual skill, mainly because the acoustic speech signal is highly variable due to various between as well as within-speaker differences. The present work is concerned with the question how the perceptual system deals with this variable nature of speech so as to interpret it correctly. The thesis consists of two parts. In the first part, we explored selective speech adaptation and visual recalibration of auditory speech. These two aftereffects dissociated in a number of ways. Moreover, we observed that the two aftereffects can occur simultaneously. In the second part of this thesis we explored whether lipread and lexical information produce comparable recalibration effects and direct bias effects on phoneme perception. The present chapter provides a summary and discussion of the main observations.

Visual recalibration versus selective adaptation

The research presented in the present thesis started with the findings by Bertelson et al. (2003). They demonstrated recalibration of speech sounds and selective speech adaptation within the same experimental setting just by manipulating the ambiguity of the auditory speech tokens. In the first chapters these two aftereffects were further investigated. Besides the consistently observed opposite directions in which the aftereffects develop, we reported a number of other dissociations demonstrating that the two aftereffects are based on different mechanisms.

In all the experiments described in the first chapters (chapters 2, 3, 4, 5, and 6), we exposed participants to audiovisual speech stimuli after which we measured for possible aftereffects on auditory-only identification trails. Participants were exposed to audiovisual congruent syllables /aba/ and /ada/ (AbVb and AdVd) in order to investigate selective adaptation to speech. Visual recalibration of auditory phoneme perception was investigated by exposing listeners to audiovisual ambiguous speech tokens, which contained an ambiguous auditory phoneme between /aba/ and /ada/, which was dubbed onto a face articulating either /aba/ (A?Vb) or /ada/ (A?Vd). During the subsequent posttests participants were asked to identify auditory ambiguous speech tokens of the /aba/-/ada/ continuum as either /aba/ or /ada/. Recalibration reveals itself as an increase in adapter consistent responses on the posttest whereas selective speech adaptation reveals itself as a decrease in adapter consistent responses on subsequent posttests.

The role of visual and auditory speech information

The experiments presented in chapter 2 were designed to investigate the role of auditory and visual information for selective speech adaptation and recalibration. A main question was why in previous studies exposure to McGurk pairs with non-ambiguous auditory tokens combined with incongruent visual tokens did not produce bigger

aftereffects than exposure to unimodal auditory tokens (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). That is, exposure to a McGurk stimulus might produce both selective speech adaptation and recalibration. Selective speech adaptation might be caused by the repeated presentation to a clear auditory phoneme and recalibration of phoneme perception which might be produced by exposure to the conflict between auditory and visual speech information. If the effects of selective speech adaptation and recalibration would sum, the aftereffects produced by exposure to the audiovisual incongruent stimuli would be bigger than the after-effects produced by exposure to audiovisual congruent, auditory-only or audiovisual ambiguous stimuli. To test this hypothesis, participants were exposed to audiovisual incongruent stimuli, i.e. auditory /aba/ dubbed onto visual /ada/ (AbVd) and auditory /ada/ dubbed onto visual /aba/ (AdVb), auditory-only /aba/ and /ada/ (Ab and Ad)and visual-only /aba/ and /ada/ (Vb an Vd) stimuli in addition to the audiovisual congruent (AbVb and AdVd) and audiovisual ambiguous (A?Vb and A?Vd) stimuli. To further investigate for dissociations between recalibration and selective speech adaptation, the stimuli were presented either 8 or 50 times subsequent to posttesting.

Results indicated that the combination of an auditory ambiguous speech sound with disambiguating lipread speech is essential for visual recalibration to occur. However, visual information had no effect on aftereffects when listeners were presented visual-only or audiovisual incongruent stimuli containing a clear auditory speech token (i.e. a McGurk-stimulus). This absence of visual effects in selective speech adaptation can therefore not be attributed to the idea that visual information as such is impotent, because visual information had a substantial impact on aftereffects when the sound was ambiguous. Visual information thus affects aftereffects, but only if combined with an ambiguous speech sound. Furthermore, the magnitude of recalibration was not affected by the number of adapter presentations.

The repeated exposure to a stimulus containing an auditory clear speech token always produced selective speech adaptation. Equal amounts of selective speech adaptation were obtained irrespective of whether the visual information was absent, congruent or incongruent to the auditory information. Furthermore, the magnitude of selective speech adaptation increased when the number of adapter presentations increased.

Time course

One of the basic properties of an aftereffect is how it builds up and dissipates over time. To further dissociate the processes underlying selective speech adaptation and recalibration we investigated dissipation (chapters 3 and 4) and build-up functions (chapter 5) of the two aftereffects.

In order to measure dissipation, we prolonged posttests subsequent to audiovisual congruent and audiovisual ambiguous exposure. Rate of dissipation of the aftereffects was thus measured as a function of post-test duration. The dissipation rates clearly dissociated selective speech adaptation and recalibration: whereas selective speech adaptation remained stable with prolonged post-testing, recalibration dissipated fast when more testtokens were presented. In the experiment described in chapter 4, participants were presented to 8 or 32 audiovisual adapter presentations subsequent to posttesting, in order to investigate for an affect of exposure duration on the dissipation rates of the two aftereffects. For recalibration, exposure duration did not affect the magnitude of the effect or the rate of dissipation. Again, the magnitude of selective speech adaptation increased when more audiovisual congruent adapters were presented and the effect did not diminish with prolonged post testing.

Recalibration effects were either not (chapter 2) or negatively (chapter 4) affected by increasing the number of exposures. This was a rather unexpected observation, although no data regarding the build-up course of recalibration were available thus far. The study presented in chapter 5 was designed to investigate the acquisition of selective speech adaptation and recalibration. The cumulative number of audiovisual congruent or ambiguous adapters was increased from 1 to 256, with posttests interspersed after 1, 2, 4, 8, 16, 32, 64, 128 and 256 adapter repetitions.

Selective speech adaptation and recalibration followed different build-up courses. As predicted, selective speech adaptation produced by the audiovisual congruent adapters increased in a log-linear way when exposure was prolonged. Thus, the more audiovisual clear adapters were presented, the stronger the effect became. For the aftereffects of the audiovisual ambiguous adapters another pattern emerged. Recalibration followed a curvilinear trend, with a fast initial build up, then reaching a plateau on which the magnitude of the effect remained stable, which was then followed by a gradual decline of the aftereffect when more adapters were presented.

The decline following prolonged exposure to the audiovisual ambiguous adapters was a surprising result. We proposed that the audiovisual ambiguous adapters may also induce selective speech adaptation causing a gradual decline of the aftereffect. The conflict between auditory and visual speech information thus induced recalibration, while the basis for selective speech adaptation was provided by the auditory stimulus being shifted towards one of the phoneme categories. The audiovisual ambiguous stimulus thus induced both recalibration and selective adaptation. Given that recalibration builds up fast and selective adaptation needs time to build up, aftereffects on the first testtrials mainly reflect recalibration, while selective speech adaptation comes into play when more adapters are presented.

Developmental trend

In chapter 6, we explored developmental trends of selective speech adaptation and recalibration. Children in two age groups, 8-year-olds and 5-year-olds were exposed to the audiovisual congruent (AbVb and AdVd) and audiovisual ambiguous (A?Vb and A?Vd) stimuli and were subsequently asked to identify tokens of the /aba/ -/ada/ phoneme continuum. For 8-year-olds, but not for 5-year-olds we observed aftereffects indicative of recalibration. The aftereffect caused by exposure to the audiovisual ambiguous adapters were bigger than the aftereffects produced by audiovisual congruent exposure for the 8-year-olds, indicating that they shifted their phoneme boundaries as a result of exposure to audiovisual ambiguous speech input. For the 5-year-olds there was no difference in the aftereffects produced by audiovisual ambiguous or audiovisual congruent adapters.

The difference in susceptibility to visually-driven phoneme recalibration for the two age groups is probably the result of the smaller effect that lipread information has on auditory speech perception in younger children. As observed before, the effect of lipread speech increases with increasing age and even 12-year-olds do not make use of visual information as effectively as adults do (McGurk & MacDonald, 1976).

Remarkably, there was an absence of selective speech adaptation effects in both age-groups. There are two possible explanations for this result. Firstly, since we were primarily interested in recalibration, only 8 adapters were presented during exposure. This is an optimal amount of exposure for inducing recalibration but a rather small amount of adapters to effectively produce selective speech adaptation. Hence, the number of adapter presentations might have been too small to produce selective speech adaptation. Secondly, the children might have a tendency to repeat what was heard during the exposure phase whenever they were unsure regarding the identity of the ambiguous phoneme during the post-tests. Such a strategy would conceal the aftereffects produced by selective speech adaptation. Since we used both types of audiovisual adapters, though, we were able to eliminate the effects of such response strategies by subtracting aftereffects from each other.

The literature on perceptual learning and audiovisual speech perception predicted opposite results for the development of recalibration. Whereas theories on perceptual learning predicted a decrease of recalibration with age, studies on audiovisual speech perception predicted an increase. The results of our study are consistent with the literature on audiovisual speech perception which has repeatedly demonstrated an increase in visual effectiveness on auditory speech perception with age. Importantly, this study also shows that perceptual aftereffects in younger children need to be interpreted with caution. In order to account for strategic response patterns, a baseline needs to be included.

Recalibration and selective adaptation in speech perception

Together, the studies presented in the first part of this thesis demonstrate that the perception of an ambiguous speech sound can be altered by two different phenomena: selective adaptation and recalibration. Although selective adapatation and recalibration rely on different processes in speech perception, their after-effectis can interact.

When listeners are confronted with an ambiguous speech sound, they make use of visual speech information to adaptively shift the phoneme boundary so that it is in congruence with the visual speech input. Recalibration occurs fast and also dissipates fast with posttesting (chapter 3, 4 and 5). Adjustment to unusual sounding speech is thus acquired rapidly and remains flexible. Under normal circumstances, such a strategy is beneficial since listeners need to quickly adapt to the different speech sounds produced by various speakers.

Speech perception is thus a dynamic process in which phoneme representations are retuned to align with the incoming speech input in order to overcome the variability of the speech signal. This flexible view on speech perception is substantiated by a rapidly growing amount of research on perceptual learning or recalibration in speech perception (Bertelson et al., 2003; 2006; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006, 2007; Norris et al., 2003; van Linden & Vroomen, in press; Vroomen et al., 2007; Vroomen, 2004). This dynamic account contrasts with a more static view on speech perception where it is hypothesised that perceptual coherence is obtained by filtering or normalization of the signal into an abstract representation (Blumstein & Stevens, 1980; Tenpenny, 1995). According to these abstractionist accounts, the variability of the speech input is overcome by filtering the signal from all idiosyncratic details. The signal is supposedly reduced to an abstract or "normalized" form. This reduced signal is than mapped onto abstractly defined, stable phoneme representations. Thus, instead of a flexible system that adjusts phoneme representations to variations in the speech signal, abstractionist theories suggest that the signal is adjusted to match the phoneme representations. A problem for such an abstractionist view on speech perception is that, as yet, no complete set of invariant physical features of acoustic speech has been uncovered (Delattre et al., 1955; A. Liberman et al., 1952; Peterson & Barney, 1952). The result presented in the present thesis and reported in other studies on lexically guided perceptual learning in speech perception (Eisner, 2006; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006, 2007; Norris et al., 2003) therefore suggests a dynamic rather than a static process for auditory speech perception.

The research on selective speech adaptation emerged from an abstractionist theory of speech perception, as it aimed to demonstrate the existence of feature specific representations of speech sounds. The research presented here, demonstrated that although originating from different views on speech perception, selective speech

adaptation and recalibration paradigms both add to the understanding of the speech perception process. Selective speech adaptation and recalibration were dissociated in the direction of their aftereffects, in their dissipation rate and in their build-up functions; implying that the effects are based on different mechanisms. That is, whereas recalibration reflects the adaptive nature of the speech perception process, selective speech adaptation probably reflects the fatigue of the relevant speech perception processes (Eimas & Corbit, 1973). The observed build-up trend for selective speech adaptation adds to the idea that selective speech adaptation relies on processes of fatigue, as the effect emerged only after a large number of adapter presentations and increased when more adapters were presented; a pattern consistent with the occurrence of neural fatigue (Samuel, 1986; Sekuler & Pantle, 1967).

However, although the effects are caused by different processes, the aftereffects of selective speech adaptation and recalibration can add-up as exposure to an audiovisual speech token containing an auditory ambiguous phoneme can produce both recalibration and selective speech adaptation. This phenomenon also explained an apparent contradiction in the literature on speech perception. That is, two studies investigating the aftereffects of exposure to an ambiguous phoneme embedded in a lexical context, reported aftereffects in opposite directions. Samuel, (2001) reported selective speech adaptation to occur after exposing participants to an ambiguous fricative between /s/ and /š/ which was presented after words normally ending in /s/ or /š/ (e.g. "abolish" and "bronchitis"). This result contrasted with the results obtained by Norris and colleagues (Norris et al., 2003), who exposed listeners to a lexically embedded ambiguous fricative between /f/ and /s/ in Dutch words like "witlof" (*chicory*) and "naaldbos" (pine forest), and reported aftereffects indicative of recalibration. The most important difference between these two studies was the number adapter presentations subsequent to posttesting. Norris and colleagues (2003) presented participants to only 20 words ending in an ambiguous sounding phoneme. In contrast, Samuel (2001) presented listeners to 24 adapter blocks, which each consisted of 32 adapter presentations. The obtained overall negative aftereffect by Samuel was calculated as the average aftereffect over all the exposure blocks. Our re-analyses of the aftereffects reported by Samuel as a function of the adapter blocks, revealed an aftereffect which developed in a similar trend as observed in our study. That is, a positive aftereffect, indicative of recalibration was present after the first adaptation block (i.e. after 32 adapter presentations). The aftereffect then became negative (i.e. indicative of selective speech adaptation) only after block 3. It is thus likely that in the study by Samuel, as in our study, exposure to ambiguous sounding speech first produced recalibration and was later overwhelmed by selective speech adaptation.

Selective speech adaptation by a restored phoneme is however not always preceded by recalibration. The adapter stimuli used by Samuel (1997) to investigate

aftereffects caused by a restored phoneme only produced selective speech adaptation with no sign of recalibration after trial-by-trial re-analyses. (Samuel 1997; Kraljic & Samuel, 2005). As the adapters were presented only 40 times subsequent to posttesting, it is not likely that no recalibration was observed because it was overwhelmed by selective speech adaptation. That is, the declination of the recalibration effect as described in chapter 5, which was most probably caused by processes of selective speech adaptation, started after 64 adapter presentations. Critically, in this study, the unusual sound on which lexical information exerted its influence was not an ambiguous phoneme, but a noise sound. As lexically driven recalibration reflects the adjustment of a phoneme boundary, it is unlikely that such adaptive processes takes place for a noise sound. Most likely, exposure to a restored phoneme produces selective speech adaptation as listeners repeatedly heard a clear phoneme.

The results by Samuel (1997, 2001) and the results presented here in chapter 5 are furthermore relevant to the debate concerning the phonetic versus the acoustic nature of selective speech adaptation as they demonstrate that ambiguous speech sounds can produce selective speech adaptation when disambiguated by lipread or lexical information. Previous studies did not find an effect of lipread information on speech aftereffects, probably because the auditory component of the adapting stimulus was a clear phoneme (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). Aftereffects driven by visual speech are only observed when the lipread speech is combined with auditory ambiguous speech. Hence, exposure to McGurk stimuli only produces selective speech adaptation since it contains an auditory clear speech sound. When the auditory speech input is unambiguous, visual speech information does not affect aftereffects.

Lipread and Lexical effects on phoneme perception

The second part of the thesis was concerned with the direct bias and aftereffects of lipread and lexical information on auditory phoneme perception. To effectively compare recalibration driven by lipread and lexical information, we measured their effects on the same ambiguous phoneme, halfway between /p/ and /t/. For the lexical stimuli, this ambiguous phoneme was presented after Dutch words which normally end in /p/ or /t/, for example the words "vloot" and "hoop" (meaning *fleet* and *hope* in English respectively). For the lipread stimuli the same ambiguous phoneme was presented after pseudowords, which were dubbed onto a face articulating that pseudoword ending in a /p/ or /t/ (e.g. auditory ambiguous woo? was dubbed onto a face articulating either woop or woot). In chapter 7, the magnitude, dissipation rate and stability over time of lipread and lexically-driven recalibration were investigated. In chapter 8 an electrophysiological study was

presented investigating the perceptual nature of lexically-driven recalibration on phoneme perception.

Visual versus lexically-driven recalibration

Lipread and lexical information are intrinsically different in nature, as lipread information is a bottom-up source of information whereas lexical information is a top-down information source. Lipread speech is thus an inherent property of the speech input, while lexical information depends on stored language-specific lexical knowledge of the listener. This difference might cause the two information sources to produce different aftereffects on phoneme perception. In line with this view, we found that whereas lipread-driven recalibration dissipates fast with prolonged posttesting, (Vroomen, et al. 2004), studies on lexically-driven recalibration reported long-lasting aftereffects (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). Here, we investigated possible explanations for these contradicting results. These may be the result of the intrinsic difference between the two sources of in disambiguating speech information, or they might have caused by the differences in stimuli and procedures between the studies investigating for either lipread or lexically driven recalibration.

All the experiments presented in chapter 7 consisted of an exposure-posttest design. During exposure, participants were presented to the ambiguous phoneme which was disambiguated by either a lipread or lexical information. Testing for aftereffects of this exposure, occurred on a subsequent auditory-only phoneme identification task in which we presented midpoint tokens of the /t/ - /p/ continuum. Participants were asked to identify these ambiguous phonemes as either /t/ or /p/.

The first experiment confirmed that lipread and lexical information both produced perceptual recalibration effects of the ambiguous phoneme when using the same stimuli and procedures. Both aftereffects were equally large in magnitude.

In the second experiment posttesting was prolonged in order to investigate the rate of dissipation of the two aftereffects. Both the lipread and lexical exposure stimuli again produced recalibration effects, with lipread stimuli producing somewhat bigger aftereffects than lexical stimuli. Importantly, as we observed before, both lipread and lexical recalibration dissipated fast with prolonged posttesting. There was thus no sign that lexical recalibration would last longer than lipread recalibration as might be expected when comparing previous reports on the stability of visual and lexical recalibration.

So far, all previous studies on lexical recalibration presented so-called contrast stimuli (i.e. words containing the clearly spoken phoneme of the other side of the continuum) in addition to the ambiguously ending words during the exposure phase. For example, the study by Norris et al. investigated lexically driven recalibration on a ambiguous fricative between /f/ and /s/. During an exposure-phase in which the ambiguous sound was presented after words normally ending in /s/, participants were also presented with words ending in a clear sounding /f/. Due to selective speech adaptation induced by the contrast phoneme or other criterion-setting operations, it might be possible

that the resulting aftereffects were either bigger or more stable over time. In Experiment 3 of chapter 6, we investigated this possibility by adding contrast stimuli during the exposure phase. During exposure to lipread or lexical information which biased perception of the ambiguous phoneme towards hearing /t/, participants were now also presented with words or audiovisual pseudowords, ending in a clear /p/. During the /p/-biasing conditions, we also included either words or pseudowords ending in a clear sounding /t/. As predicted, this resulted in bigger aftereffects for both lipread and the lexical exposure. The additional contrast stimuli thus indeed enlarged the perceptual shift on the posttests. The aftereffects produced by the lipread stimuli were also again bigger in magnitude than the aftereffects produced by the lexical stimuli. However, although the presentation of contrast stimuli increased the magnitude of the aftereffects, they did not become more stable over time. Both effects thus again dissipated fast during posttesting.

Previous studies investigating the stability over time of lexical recalibration introduced a delay between exposure and posttests (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). In contrast, we tested dissipation as a function of the position of the test token in the posttest that was presented immediately after exposure. This difference in procedure raised the possibility that the fast dissipation in our studies was caused by the presentation of the ambiguous sounding posttest trials immediately after exposure. Experiment 4 controlled for this possibility as we introduced a 3 minute silence interval between the exposure phase and the posttest. Aftereffects were still observed after this silence interval and were, on the first testtokens positions, of the same size as those observed in experiment 2. Lipread and lexical aftereffects thus remained stable until the posttest phase started, and then quickly dissipated. If the mere passing of time diminishes recalibration, no aftereffects would be observable after this time period as in the previous experiments the aftereffects had gone after about 1 minute.

A final experiment investigated whether the use of a between-subjects design would produce more stable aftereffects. In most of our experiments, participants were exposed to both types of exposure stimuli. As a consequence, the phoneme boundary of the participant thus continuously shifted from one side of the continuum to the other. This might negatively affect the stability of recalibration. Therefore, we now recalibrated the phoneme boundary towards only one side of the continuum. However, results showed no sign that aftereffects were bigger or would last longer if participants were exposed to one instead of both phoneme categories.

In addition to the exploration of lipread and lexical recalibration, we also studied whether the immediate bias effects phoneme perception. As predicted, lipread information produced bigger direct bias effects than lexical information. For the lexical stimuli, we also observed a positive relation between the size of immediate bias and recalibration. This relation is noteworthy when considering the level at which lexical information exerts its

influence on speech processing. As mentioned, a direct bias on phoneme perception produced by lexical context is sometimes argued to reflect on phonemic decision level. Recalibration, on the other hand, supposedly reflects a shift of the phonetic boundary at a perceptual level. The correlation between direct bias and the size of the aftereffect might suggest, though, that the processes underlying direct bias and recalibration are the same.

The perceptual nature of lexical recalibration

The electrophysiological study presented in chapter 8 explored the extent to which lexically driven recalibration of phoneme perception is perceptual in nature. The notion is that lexical recalibration induces a shift in the boundary between two phonetic categories, and to the extent that this shift occurs at a perceptual level, one may observe that lexical information actually affected early processing stages. This prediction was tested using recordings of human event-related brain potentials (ERPs) focusing on the mismatch negativity (MMN). The MMN is an pre-attentive, auditory evoked brain potential, elicited by a discriminable change in an otherwise constant acoustic or phonological stimulus train (Näätänen et al., 1978).

In an odd-ball paradigm, Dutch listeners were presented to a word normally ending in /t/ (vloot, meaning 'fleet' in English) or in /p/ (hoop, meaning 'hope'), whereby the final consonant (/t/ or /p/) was replaced by an ambiguous sound halfway between /t/ and /p/ (henceforth /?/). The non-ambiguous sound /t/ embedded in the same context served as the deviant stimulus in both conditions. Importantly, the physical difference of the final phonemes between the standard (i.e. the ambiguous phoneme /?/) and deviant stimulus (i.e. clear /t/) was thus the same for both conditions. Perceptually however, the final phoneme of the standard stimulus in the "hoop" condition was recalibrated towards hearing /p/, whereas the final phoneme of the standard stimulus during the "vloot" condition was recalibrated towards hearing /t/. We investigated whether this perceptual difference was reflected on the MMN.

It was indeed observed that lexical information produced a change the amplitude of the MMN-response. The amplitude of the MMN-response evoked by the difference between the ambiguous sound and unambiguous /t/ was bigger for the p-word 'hoop' than the t-word 'vloot'. This result thus implies that lexical information reached down to early prelexical processing stages. This observation strengthens the view that the lexically driven shift in phoneme perception is indeed perceptual in nature.

Lipread and lexical effects on speech perception

The research presented in the second part of the thesis demonstrated that, from a functional perspective, there is no difference between bottom-up perceptual information

and top-down stored lexical knowledge for speech perception. Listeners use both information sources in the same way to correctly interpret ambiguous speech sounds in online perception and to adjust the boundary between two phonetic categories. It is therefore concluded that lipread and lexical information serve a similar role in the perception of auditory speech: maintaining perceptual coherence of speech input. The recalibration produced by lipread and lexical information rely on similar mechanisms as their aftereffects not only dissipate in similar trends, but are also similarly affected by the various procedural variations.

Recalibration processes occur fast and relatively automatically, as in all our experiment listeners were presented to only 8 auditory ambiguous stimuli and were simply instructed to passively observe the stimuli. This distinguishes recalibration from other forms of perceptual learning in speech perception, which require long periods of learning and involve explicit discrimination instructions (Logan et al., 1991). Of course, the two types of learning are functionally different as the latter type of perceptual learning involves the acquirement of a new phoneme contrast, while recalibration as described in the present thesis reflects a shift in the perception of an already existing continuum in order to adjust perception to ambiguous speech sounds. Adjusting an existing phoneme boundary probably is less arduous and also needs to occur faster and more automatic than learning an entirely new phoneme.

Recalibration of phonemic identity could be the result of several perceptual changes. It could for example, involve a change in the position of the phoneme prototype, or it could represent the adding of the ambiguous phoneme as a new exemplar of the particular category. Most probably, recalibration represents a shift in the phoneme boundary of the particular phoneme contrast. If recalibration would involve the moving of a phoneme prototype, one would expect to find perceptual changes on this category prototype. The results obtained in chapter 2 do not support such a mechanism for recalibration, as we did not find any perceptual shifts on the continuum endpoint tokens after exposure to the audiovisual ambiguous stimuli. Effects of recalibration were only observed on the ambiguous midpoint tokens of the phoneme continuum. Second, exposure to an ambiguous speech sound with disambiguating lipread or lexical information probably does not merely add this sound as a new exemplar of the specific category. If so, recalibration would occur only on the specific phoneme which was present during exposure. This is not what we observed, as perception was not only shifted for this phoneme: we consistently observed that perception of the two neighbor-tokens on the continuum was also shifted. Norris et al. (2003) and Kraljic and Samuel (2005) also observed that the perceptual shift is not limited to the ambiguous sound presented during exposure. In addition, if the ambiguous sound would be fully incorporated into the particular phoneme category, recalibration effects would be complete. Recalibration-effects

were however not complete relative to the direct bias effect (Bertelson et al., 2003). In general, shifts in perception due to recalibration are about a quarter of the imposed discrepancy (Epstein, 1975; Welch, 1978a, 1986). Such a pattern is indeed the predicted when recalibration represents a shift in the phoneme boundary.

The research presented here, demonstrated that there is a flow of information from lexical to prelexical levels and from areas processing visual speech to a prelexical stage. The results presented in chapter 8 demonstrated that the shift in phoneme perception indeed occurs at an early prelexical level, strengthening accounts that acknowledge the perceptual nature of the phenomenon (Bedford, 1999; Bertelson, 1998; McQueen, 2003; Vroomen & de Gelder, 2004a). Interactive models on speech processing provide a straightforward explanation for this flow of information as they assume feedback projections from the lexical onto the prelexical level. Both the online effect and the recalibration effects can be explained in this manner. TRACE (McClelland & Elman, 1986) accounts for effects of perceptual learning as the network of interactive connections between prelexical and lexical layers updates itself by strengthening or retuning the connections between units which were simultaneously activated. Lexical information can reach down and retune prelexical mechanisms, thereby mediating boundary adjustment.

Autonomous approaches on speech perception, such as the Merge model (Norris et al., 2000, 2003), state that the direct effects of lexical information arise on a decision level whereas lexically-driven recalibration is caused by a specialized off-line feedback mechanism affecting speech processing over a longer time period (Norris et al. 2003; McQueen 2003). In the Merge model it is proposed that the lexical level provides a training signal that can be used to alter the way prelexical information is interpreted. This offline feedback for learning takes place without the necessity for online feedback from lexical onto pre-lexical levels of processing. As described in Norris et al. (2003), this training signal driving feedback for learning could be carried by a separate pathway, thereby not affecting online processing. As an alternative, the authors proposed that the training signal might be constructed from the same neural components as the feedforward pathways. As a consequence, information provided by the lexical level could be available to online processing mechanisms. Feedback for online processing would than only exits as an epiphenomenon of a mechanism which is required for perceptual learning.

At present, most theories on speech perception focus on either lipread or lexical effects on auditory speech perception (Liberman & Mattingley, 1985; Massaro, 1987; McClelland & Elman, 1986; Norris et al., 2000; Summerfield, 1987). A comprehensive model on speech processing should however incorporate both as they both affect auditory speech perception. As was also stated by Ernst & Banks (2002; 2004), the brain probably takes advantage of the strength of each modality, such that the information that is the most accurate encoded in one modality influences perception in the other modality.

The research presented in the present thesis, suggests that the mechanism to maintain perceptual coherence in speech perception, corresponds to the mechanism and function of recalibration by pairing which has been demonstrated for many other perceptual domains (Bedford, 1999; de Gelder & Bertelson, 2004; Epstein, 1975; Vroomen & de Gelder, 2004a). Recalibration by pairing occurs under conditions of exposure to a conflict situation where discrepant information input is provided regarding a specific characteristic of a stimulus. The information in the modalities is than changed, i.e. recalibrated, such that the perceived discrepancy is reduced. This is exactly what occurs for both lipread and lexically-driven recalibration of phoneme perception. Due to the registered discrepancy between auditory and lipread speech or between auditory speech and lexical knowledge, information processing for auditory phoneme perception is altered so as to maintain perceptual coherence. Instead of a proposal for different processes underlying online- and recalibration effects (i.e. phoneme-decision and off-line feedback for learning) it might be more appropriate to incorporate both effects into a single mechanism which is responsible for the maintenance of perceptual coherence in speech perception. One plausible mechanism is that listeners flexibly adjust their phoneme boundary whenever an ambiguous sound is heard in a disambiguating context. Bias on phoneme categorization and aftereffects are, on this view, thus caused by the same mechanism. A testable prediction for such a mechanism would be that recalibration only occurs when there is a bias.

References

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

Beachamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information in the superior temporal sulcus. *Neuron, 41,* 809-823.

Bedford, F. L. (1989). Constraints on learning new mappings between perceptual dimensions. *Journal of Experimental Psychology-Human Perception and Performance, 15*, 232-248.

Bedford, F. L. (1999). Keeping perception accurate. *Trends in Cognitive Sciences, 3,* 4-11.

Bell-Berti, F., & Harris, K. S. (1981). A temporal modal of speech production. *Phonetica, 38*, 9-20.

Bertelson, P. (1996). Starting from the ventriloquist: The perception of multimodal events. *International Journal of Psychology, 31*, 4291-4291.

Bertelson, P. (1998). Starting from the ventriloquist: The perception of multimodal events. In M. Sabourin, F. Craik & M. Robert (Eds.), *Advances in psychological science, vol 2: Biological and cognitive aspects* (pp. 419-439). Hove: Psychology Press.

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

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). *Exploring the relation between mcgurk interference and ventriloquism.* Paper presented at the International Congress on Spoken Language Processing, Yokohama.

Bertelson (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersleben, T. Bachmann & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347-363). North-Holland: Elsevier.

Bertelson, P., & De Gelder, B. (Eds.), (2003). *The psychology of multimodal perception*: Oxford University Press.

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

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

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67,* 648-662.

Boersma, P., & Weenink, D. (2002). Praat: Doing phonetics by computer (Version 4.0.7). Amsterdam.

Bradlow, A. R., Pisoni, D. B., Akehane-Yamada, R., & Tohkura, Y. (1997). Training japanese listeners to identify enlish /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101,* 2299-2310.

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

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*, 204-220.

Burnham, D., & Sekiyama, K. (2004). When auditory-visual speech perception develops: The locus of the Japanese McGurk effect. *Australian Journal of Psychology, 56*, 108-108.

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

Calvert, G., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes*. Cambridge, MA: The MIT Press.

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*, 649-657.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., P., M. and Suckling, J. (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). *Cognitive Brain Reseach, 12*, 233-243.

Choe, S. C., Welch, R. B., Gilford, R. M., Juola, J. F. (1975). The "ventriloquisteffect" Visual dominance or response bias? *Perception and Psychophysics 18,* 55-60.

Clarke, C. M. (2000). Perceptual adjustment to forgeign accented english. *Journal of the Acoustical Society of America, 107*, 2856.

Colin, C., Radeau, M., Deltenre, P., & Morais, J. (2001). Rules of intersensory integration in spatial scene analysis and speechreading. *Psychologica Belgica, 41*, 131-144.

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology, 115*, 1989-2000.

Connine, M. C., Titone, D., Deelman, T., & Blasko, D. (1997). Similarity mapping in spoken word recognition. *Journal of Memory & Language, 37,* 463-480.

Cooper, F. S. (1974). Contingent feature analysis in speech perception. *Perception & Psychophysics, 16,* 201-204.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987a). Phoneme identification and the lexicon. *Cognitive Psychology, 19,* 141-177.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*, 460-467.

de Gelder, B., & Bertelson, P. (2004). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 8*, 7-7.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport, 8*, 919-924.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consontants. *Journal of the Acoustical Society of America, 27*, 769-773.

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

Dupoux, E., & Green, K. P. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception & Performance, 23*, 914-927.

Dupoux, E., & Mehler, J. (1990). Monitoring the lexicon with normal and compressed speec: Frequency effects and the pre-lexical code. *Journal of Memory & Language, 29*, 316-335.

Eimas, P. D., Cooper, W. E., & Corbit, J. D. (1973). Some properties of linguistic feature detectors. *Perception and Psychophysics, 13*, 247-252.

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

Eisner, F. (2006). *Lexically-guided perceptual learning in speech processing.* Radboud Universiteit Nijmegen, Nijmegen.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics, 67*, 224-238.
Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America, 119*, 1950-1953.
Epstein, W. (1975). Recalibration by pairing: A process of perceptual learning. *Perception, 4*, 59-72.
Ernst, M.O. and Bülthoff H.H., (2004) Merging the Senses into a Robust Percept. *Trends in Cognitive Sciences, 8*, 162-169.
Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429-433.
Fernandez-Ruiz, J., & Diaz, R. (1999). Prism adaptation and aftereffect: Specifyingthe properties of a procedural memory system. *Learning & Memory, 6*, 47-53.
Fox, R. A. (1984). Effects of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception & Performance, 10*, 526-540.
Frissen, I. (2005). *Visual recalibration of auditory spatial perception.* Tilburg University, Tilburg.
Frissen, I., Vroomen, J., de Gelder, B., & Bertelson, P. (2003). The aftereffects of ventriloquism: Are they sound-frequency specific? *Acta Psychologica, 113*, 315-327.
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance, 6*, 110-125.
Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology, 19*, 29-56.
Gilbert, C. D. (1994). Early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America, 91*, 1195-1197.
Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron, 3*, 681-697.
Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*, 2677-2690.
Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr Clin Neurophysiol, 55*, 468-484.
Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories on speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.). *Hearing by eye ii: Advances in the psychology of speechreading and auditory-visual speech.* Hove, England: Psychology Press.
Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*, 524-536.
Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14*, 421-433.
Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6-12. *Journal of Phonetics, 28*, 377-396.
Held, R. (1965). Plasticity in sensory-motor systems. *Scientific American, 213*, 84-94.
Hershenson, M. (1989). Duration, time constant and decay of the linear motion aftereffects as a function of inspection duration. *Perception and Psychophysics, 45*, 251-257.
Hockley, N. S., & Polka, L. A. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America, 96*, 3309.

Jordan, T., & Sergeant, P. (1998). Effects of facial image size on visual and audiovisual speech recognition. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 155-176). London: Psychology Press.

Jordan, T., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43,* 107-124.

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

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

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

Kuhl, P. K. (2000). A new view of language acquisition. *Proc Natl Acad Sci U S A, 97*, 11850-11857.

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

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior & Development, 7*, 361-381.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental chance. *Journal of the Acoustical Society of America, 100*, 425-438.

Ladenfoged, P, & Broadbent, D. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98-104.

Lang, H., Nyrke, T., Ek, M., Aaltonen, O., Raimo, I., & Näätanen, R. (1990). Pitch discrimination performance and auditory event-related potentials. In C. H. M. Brunia, A. W. K. Gaillard, A. Kok, G. Mulder & M. N. Verbaten (Eds.), *Psychophysiological brain research* (Vol. 1, pp. 294-298). Tilburg, The Netherlands: Tilburg University Press.

Lenneberg, E. (1967). *Biological foundations of language.* New York: Wiley.

Liberman, A., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of unvoiced stop-consonants. *Journal of Psychology, 65*, 497-516.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology, 52*, 127-137.

Liberman, A. M., & Mattingley, J. G. (1985). The motor theory of speech perception revised. *Cognitive Psychology, 21*, 1-36.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify english /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89,* 874-886.

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.

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

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

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

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences, 10*, 533.
Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25*, 457-465.
Montgomery, A. A., Walden, B. E., & Prosek, R. A. (1987). Effects of consonatal context on vowel lipreading. *Journal of Speech and Hearing Research, 30*, 50-59.
Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research, 17*, 154-163.
Morrongiello, B. A., Robson, R. C., Best, V. C. T., & Clifton, R. K. (1984). Trading relations in the perception of speech by 5-year-old children. *Journal of Experimental Child Psychology, 37*, 231-250.
Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13*, 417-425.
Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*, 351-362.
Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. Head movements improves auditory speech perception. *Psychological Science, 15*, 133-137.
Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America, 104*, 530-539.
Näätänen, R. (1992). *Attention and brain function*. Hilsdale.
Näätänen, R. (1999). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology, 38*, 1-21.
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., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Viano, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*, 432-434.
Näätänen, R., Paavilainen, P., Tiitinen, H., Jiang, D., & Alho, K. (1993). Attention and mismatch negativity. *Psychophysiology, 30*, 436-450.
Nittruoer, S., & Studder- Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research, 30*, 319-329.
Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299-370.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204-238.
Ohde, R. N., & Sharf, D. J. (1979). Relationship between adaptation and the percept and transformations of stop consonant voicing: Effects of the number of repetitions and intensity of adaptors. *Journal of the Acoustical Society of America, 66*, 30-45.
Paré, M., Richler, R., C., ten Hove, M., & Munhall, K., G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics, 65*, 553-567.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184.
Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 1037-1052.
Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., et al. (2001). Memory traces for words as revealed by the mismatch negativity. *NeuroImage, 14*, 607-616.
Radeau, M. (1994). Ventriloquism against audio-visual speech: Or, where Japanese-speaking barn owls might help. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 13*, 124-140.
Radeau, M., & Bertelson, P. (1974). The after-effects of ventriloquism. *The Quarterly Journal of Experimental Psychology, 26*, 63-71.
Recanzone, G. H. (1998). Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 869-875.
Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics, 30*, 309-314.
Rosen, S. (1992). Temporal information in speech: Acoustic auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 336*, 367-373.
Rosenblum, L. D. (1994). How special is audiovisual speech integration? *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 13*, 110-116.
Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*, 347-357.
Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken non-words. *Perception & Psychophysics, 19*, 394-398.
Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analyses of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia, 45,* 587-597.
Saldana, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics, 54*, 406-416.
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,* 3658-3661.
Sams, M., Aulanko, R., Hamalainen, 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*, 141-145.
Samuel, A. G. (1981a). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474-494.
Samuel, A. G. (1981b). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1123-1131.
Samuel, A. G. (1986). Red herring detectors and speech perception: In defence of selective adaptation. *Cognitive Psychology, 18*, 452-499.
Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*, 97-127.

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

Samuel, A. G., & Newport, E. L. (1979). Adaptation of speech by non-speech: Evidence for complex acoustic cue detectors. *Journal of Experimental Psychology: Human Perception & Performance, 5*, 563-578.

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

Sawusch, J. R., Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 408-421.

Schwartz, J. L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye ii: Advances in the psychology of speechreading and auditory-visual speech*. Hove, England: Psychology Press.

Sekiyama, K., & Burnham, D. (2004). *Issues in the development of auditory-visual speech perception: Adults, infants, and children.* Paper presented at the 8th International Conference on Spoken Language Processing, Jeiu Island, Korea.

Sekiyama, K., Burnham, D., Tam, H., & Erdener, D. (2003a). *Auditory-visual speech perception development in Japanese and English speakers.* Paper presented at the Auditory-Visual Speech Processing, st. Jorioz, France.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003b). Auditory-visual speech perception examined by fMRIi and PET. *Neuroscience Research, 47*, 277-287.

Sekuler, R., & Pantle, A. (1967). A model for after-effects of seen movement. *Vision Research, 7*, 427-439.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception & Performance, 28,* 1447-1469.

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

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

Summerfield, Q. (1975). A*coustic and phonetic components of the influence of voice changes and identification times for CVC syllables*. The Queen's University of Belfast, Belfast, Ireland.

Summerfield, A. Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 335*, 7-71.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 3-51). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sussman, J. E. (1993). Auditory processing in children's speech perception: Results of selective adaptation and discrimination tasks. *Journal of Speech and Hearing Research, 36*, 380-395.

Sussman, J. E., & Carney, A. E. (1989). Effects of transition length on the perception of stop consonants by children and adults. *Journal of Speech and Hearing Research, 32*, 151-160.

Sussman, E., Winkler, I., Huotilainen, M., Ritter, W., Näätänen, R. (2002).Top-down effects can modify the initially stimulus-driven auditory organization, Brain Research. *Cognitive Brain Research, 13,* 393-405

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 2*, 339-363.
Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*, 475-472.
van Linden, S., & Vroomen, J. (in press). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception & Performance, x*, xxx-xxx.
van Linden, S., Stekelenburg, J. J., Tuomainen, J., & Vroomen, J. (in press). Lexical effects on auditory speech perception: An electrophysiological study. *Neuroscience Letters, x*, xxx-xxx
van Linden S., & Vroomen, J., (in press). Recalibration and selective speech adaptation in children. *Journal of Child and Language x*, xxx-xxx.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2003). *Electrophysiology of auditory-visual speech integration.* Paper presented at the Proceedings of AVSP'2003, France.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A.*
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*, 598-607.
Vatakis, A., & Spence, C. (*in press*). Crossmodal binding: Evaluating the 'unity assumption' using audiovisual speech and non-speech stimuli. *Perception & Psychophisics, x*, xxx-xxx
Vroomen, J. (1992). *Hearing voices and seeing lips: Investigations in the pschychology of lipreading.* Tilburg University, Tilburg.
Vroomen, J. (submitted). Audiovisual integration continues without attention.
Vroomen, J., & de Gelder, B. (2004a). Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.
Vroomen, J., & de Gelder, B. (2004b). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology-Human Perception and Performance, 30*, 513-518.
Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of Experimental Psychology: Human Perception & Performance, 32*, 1063-1071.
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*, 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.
Wallach, H. (1968). Informational discrepancy as a basis of perceptual adaptation. In S. J. Freedman (Ed.), *The neuropsychology of spatially oriented behavior* (pp. 209-229). Homewood, IL: The Dorsey Press.
Wallach, H., & Karsh, E. B. (1963). The modification of stereoscopic depth-perception and the kinetic depth effect. *American Journal of Psychology, 45*, 205-217.
Wallach, H., Moore, M. E., & Davidson, L. (1963). Modification of stereoscopic depth-perception. *American Journal of Psychology, 76*, 191-204.
Warren, D. H. (1970a). Intermodality interactions in spatial localization. *Cognitive Psychology, 1*, 114-133.

Warren, D. H. (1970b). Perceptual restoration of missing speech sounds. *Science, 167*, 392-393.

Welch R. B. (1978a). *Perceptual modification: Adapting to altered sensory environments*. New York, NY: Academic Press.

Welch, R. B. (1986). Adaptation of space perception. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 24(21) - 24(45). New York: Wiley & Sons.

Welch, R. B. (1972). The effect of experienced limb identity upon adaptation to simulated displacement of the visual field. *Perception and Psychophysics, 12*, 453-456.

Welch, R. B. (1978b). *Perceptual modification: Adapting to altered sensory environments*. New York, NY: Academic Press.

Welch, R. B. (1999). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In G. Aschersleben, T. Bachmann & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 371-387). Amsterdam: Elsevier.

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.

Samenvatting

Recalibratie van de auditieve foneem percepptie door liplees en lexicale informatie.

Het spraaksignaal is een zeer complex akoestisch signaal dat de meeste mensen toch zonder veel moeite begrijpen. Het verstaan en begrijpen van spraak is vooral zo moeilijk omdat het akoestische spraaksignaal erg variabel is. Deze variabiliteit wordt onder andere veroorzaakt door verschillen tussen sprekers in bijvoorbeeld stemgeluid en spreekstijl. Hetzelfde woord uitgesproken door verschillende sprekers levert daarom een geheel anders akoestisch patroon op. Wanneer het akoestische signaal van een foneem (een foneem is de kleinste klankeenheid die een verschil in de betekenis van een woord kan uitmaken), voor een luisteraar onduidelijk (ambigu) is, maken luisteraars gebruik van andere informatie bronnen om de spraak juist te interpreteren. De zichtbare lipbewegingen van de spreker (liplees informatie of visuele spraakinformatie) en kennis van woorden (lexicale informatie) worden dan gebruikt om de identiteit van de ambigue klank vast te stellen en om de toekomstige waarneming van dit foneem aan te passen. Wanneer je als luisteraar een spreker bijvoorbeeld het woord "b/d-eker" hoort zeggen, waarbij het eerste foneem wordt uitgesproken als een klank liggend tussen een /b/ en /d/, is aan de lipbewegingen te zien of de spreker "beker" of "deker" zei. Lexicale informatie zal de perceptie van de spraakklank in de richting van een /b/ trekken, omdat "beker" wel een woord is, maar "deker" niet. Hoewel het gebruik van liplees en lexicale informatie voor de directe interpretatie van spraakklanken al langer bekend is, (McGurk & MacDonald, 1978; Ganong, 1981), zijn de na-effecten in de auditieve spraakwaarneming hiervan pas recentelijk aangetoond (Bertelson, Vroomen & de Gelder, 2003; Norris, McQueen & Cutler, 2003). Deze na-effecten worden geobserveerd als een verschuiving in de waarneming van een ambigue spraakklank, als gevolg van een periode van blootstelling aan deze ambigue spraakklank in combinatie met liplees of lexicale informatie. De perceptie van de ambigue klank is verschuift dan in de richting van de liplees of lexicale informatie waarmee het voorheen gepresenteerd werd. Deze verschuiving in de waarneming van de ambigue klank wordt recalibratie genoemd. De waarneming van spraak is dus een dynamisch proces: de hersenen passen zich voortdurend aan, aan de variaties van het spraaksignaal.

In navolging van de eerste studie waarin recalibratie van auditieve spraakwaarneming door middel van liplees informatie werd aangetoond, stelden we de luisteraars in de experimenten beschreven in de eerste hoofdstukken van dit proefschrift (hoofdstukken 2, 3, 4, 5 en 6) herhaaldelijk bloot aan een ambigu spraakgeluid liggend tussen de klanken /aba/ en /ada/. Dit spraakgeluid ligt op de zogenaamde foneemgrens van /aba/ en /ada/. Dit ambigu foneem werd herhaaldelijk gepresenteerd in combinatie met een gezicht dat /aba/ of /ada/ uitsprak. Na deze presentatie, die we de exposure-fase

noemen, begon de testfase. Tijdens deze testfase hoorden de luisteraars alleen de ambigue spraakklank, nu dus zonder lipleesinformatie, en zij moesten deze klank nu beoordelen als /aba/ of /ada/. Wanneer de luisteraars de ambigue foneem tijdens de exposure-fase in combinatie met lipleesinformatie /aba/ hoorden, beoordeelden ze deze klank veelal als /aba/. Wanneer de luisteraars de ambigue spraakklank echter in combinatie met liplees informatie /ada/ hadden gehoord, beoordeelden ze dezelfde spraakklank in de testfase echter als /ada/. De foneemgrens verschoof n de richting van de visueel gepresenteerde spraak.

In de eerste hoofdstukken van dit proefschrift is dit recalibratie-effect steeds vergeleken met een ander na-effect, namelijk selectieve adaptatie. Selectieve adaptatie is een na-effect in spraakwaarneming en wordt veroorzaakt door de herhaaldelijke blootstelling aan een duidelijk klinkend spraakgeluid. Als gevolg van deze herhaaldelijke blootstelling wordt het auditieve systeem minder gevoelig voor deze klank (Eimas & Cooper, 1973). Wanneer een luisteraar herhaaldelijk de klank /aba/ heeft gehoord, zal een ambigue klank liggend op de foneemgrens van /aba/ en /ada/ daardoor als /ada/ worden waargenomen.

Hoofdstuk 2 onderzoekt de rol van visuele (liplezen) en auditieve spraakinformatie voor het optreden van recalibratie en selectieve adaptatie in spraakwaarneming. Een belangrijke vraag was of ook de perceptie van een duidelijke spraakklank gerecalibreerd kan worden, of dat recalibratie alleen plaats vindt op ambigue spraakklanken. Tevens is onderzocht of het zien van lipbewegingen in isolatie (zonder auditieve spraak) een effect heeft op de latere waarneming van een ambigu foneem en onderzochten we de rol van visuele spraakinformatie bij het optreden van selectieve adaptatie. Uit de resultaten bleek dat er geen perceptuele recalibratie plaats vindt van duidelijk uitgesproken spaakklanken. Ook veroorzaakt het zien van alleen lipbewegingen geen recalibratie. De gecombineerde presentatie van lipleesinformatie met een ambigue spraakklank is dus cruciaal is voor het optreden van recalibratie. Wat betreft selectieve adaptatie werd er geen effect gevonden van visuele spraak informatie. Wanneer de stimulus tijdens de exposure-fase een duidelijke spraakklank bevatte, was het selectieve adaptatie effect altijd van dezelfde grootte, ongeacht de aan- of afwezigheid van liplees informatie of de identiteit van de lipleesinformatie.

Het tijdsverloop van recalibratie en selectieve adaptatie is onderzocht in de hoofdstukken 3, 4, en 5. In de hoofdstukken 3 en 4 werd onderzocht hoe lang recalibratie en selectieve adaptatie aanwezig blijft, en in hoofdstuk is de opbouw van beiden effecten onderzocht.

De resultaten toonden aan dat recalibratie redelijk snel verdwijnt: kort na aanvang van de testperiode was het effect niet meer observeerbaar. Recalibratie is echter ook snel aanwezig: er treedt al na enkele aanbiedingen van de audio-visueel ambigue

spraak stimulus een perceptuele verschuiving van de ambigue klank op. Selectieve adaptatie heeft echter meer tijd nodig om op te bouwen en nam gedurende de testperiode niet af. Het effect van selectieve adaptatie bleef stabiel gedurende de testperiode. Een zeer opvallende bevinding was dat naarmate er steeds meer presentaties van de audiovisueel ambigue spaak stimulus werd aangeboden, het recalibratie-effect in eerste instantie steeds groter werd, daarna constant bleef in grootte, en het effect bij de aanbieding van nog meer presentaties juist kleiner werd. Een verklaring hiervoor is dat na langdurige blootstelling aan een audio-visuele spraakklank is, selectieve adataptatie en recalibratie gezamenlijk optreden. Het effect van recalibratie wordt als het ware overspoelt door de effecten van selectieve adaptatie, en het neemt het observeerbare recalibratie effect af.

De waarneming van spraak kan dus veranderd worden door processen van recalibratie en selectieve adaptatie. Hoewel recalbratie en selectieve adaptatie beiden waarneembaar zijn als een na-effect in spraakwaarneming, berusten ze, gezien de dissociaties in richting en het tijdsverloop van de na-effecten, alsook in de de rol van visuele en auditieve spraakinformatie, op andere processen in de spraakwarneming.

Het experiment in hoofdstuk 6 onderzocht of ook bij kinderen visuele spraak informatie de perceptie van een ambigue klank aangepast wordt en of bij hen ook effecten van selectieve adaptatie observeerbaar zijn. Bij een groep kinderen van 8 jaar vonden we inderdaad aanwijzigen van recalibratie maar bij een groep 5-jarige kinderen niet. Deze bevinding is in overeenstemming met eerdere studies waaruit bleek dat met de leeftijd, visuele spraak informatie een steeds grotere invloed heeft op de waarneming van auditieve spraak (McGurk & MacDonald, 1978). Omdat de kinderen de neiging vertoonden steeds te herhalen wat tijdens de exposure fase waargenomen werd, werd in geen van beide groepen selectieve adaptatie geobserveerd.

Met de experimenten in hoofdstuk 7 is een vergelijking gemaakt tussen visueel en lexicaal gestuurde recalibratie in spraakwaarneming. Omdat visuele spraakinformatie en lexicale informatie twee zeer verschillende informatie bronnen zijn, werd onderzocht of liplees en lexicaal gestuurde recalibratie-effecten verschillen in bijvoorbeeld extensie of tijdsverloop. Samenvattend kan geconcludeerd worden dat lipgelezen spraakinformatie en lexicale informatie eenzelfde verandering in de waarneming van een ambigu foneem teweeg brengen. De geobserveerde recalibratie-effecten waren even groot, waren even stabiel in de tijd, en reageerden hetzelfde op veranderingen in het design van het experiment.

Hoofdstuk 8 beschrijft een elektrofysiologisch experiment waarmee onderzocht werd op welk niveau van de waarneming lexicale recalibratie plaats vindt. Meer specifiek werd gezocht naar een effect van lexicale informatie op de amplitude van de mismatch negativity (MMN). De MMN is een auditief getriggerde hersenpotentiaal, dat gevoelig is

voor veranderingdingen in een geluidspatroon en al vroeg in het waanemingsproces gegenereerd wordt (Näätänen, Gaillard & Mäntysalo, 1978). Er werd inderdaad een effect van lexicale informatie op de amplitude van de MMN gevonden. Dit betekend dat, in de auditieve cortex, lexicale informatie al vroeg zijn effect heeft bij het de waarneming van spraak.

Met dit proefschrift is getracht een bijdrage te leveren aan het begrip van het proces onderliggend aan de waarneming van spraak. De resultaten van het hier gepresenteerde onderzoek laten zien dat het waarnemen van spraak een dynamisch en adaptief proces is. Wanneer we als luisteraar geconfronteerd worden met ongewoon klinkende spraakgeluiden, maken we gebruik van zowel visuele spraakinformatie, lipleesinformatie, alsook van lexicale informatie dat als kennis is opgeslagen in het brein, om de perceptie van dit spraakgeluid aan te passen. Op deze manier wordt ondanks de grote variabiliteit in het akoestische signaal, perceptuele constantheid van het spraaksignaal voor de luisteraar behouden.

Dankwoord

Jean speelt een grote rol in het verloop van het gehele traject dat ik op de Universiteit van Tilburg doorlopen heb. Ik herinner me nog goed dat ik als pas gestarte eerstejaars psychologie studente, vol goede moed naar kamer P518 ging, om daar als proefpersoon deel te nemen aan een experiment van dr. J. Vroomen. Het was het eerste onderzoek waaraan ik mee deed, in de eerste weken van mijn studie. Ik had nog nooit zoiets ongelofelijk saais meegemaakt. Piepjes en lampjes... wat had dat nou met psychologie te maken? Hier wilde ik zo weinig mogelijk mee te maken krijgen. Toch kwam ik er gedurende mijn studie achter dat er een hoop interessants zat achter de ogenschijnlijk eenvoudige en veel te langdurige taken waaraan studenten binnen de afdeling psychonomie onderworpen werden. De bestudering van de werking van het brein en zijn al functies, zoals het waarnemen, leren, aandacht en taalvaardigheden vond ik erg interessant en ik koos voor psychonomie als afstudeerrichting. In mijn laatste jaar vertelde Jean me over zijn onderzoek en bood me aan om bij hem mijn afstudeeronderzoek te doen. Bij Jean kreeg ik de smaak van het onderzoek doen te pakken en ben toen pas gaan denken om daarvan mijn werk te maken. Tot mijn geluk, was Jean op dat moment op zoek naar twee promovendi, en zo ben ik aan de UvT begonnen aan mijn promotieonderzoek. Tijdens die vier jaar heb ik op mijn beurt veel studenten lange en heel saaie taken laten doen, en ik heb eigenlijk nog nooit zoiets interessants gedaan..

Mijn grootste dank gaat dan ook uit naar Jean, die me enthousiast heeft gemaakt voor het onderzoek binnen de psychonomie en die het ontstaan van dit proefschrift mogelijk heeft gemaakt. Jean, dank je! Je had altijd tijd voor vragen of voor het uitwisselen van ideeën voor nieuwe experimenten. Van je kritische blik waarmee je designs en onderzoeksvragen in een oogwenk doorziet, heb ik veel opgestoken. Ik prijs mezelf gelukkig zo'n goede promotor te hebben gehad.

Ook Paul Bertelson wil graag bedanken. Paul, ik respecteer en waardeer je zeer. Op cruciale momenten deelde je je schijnbaar onuitputtelijke kennis of legde je vinger precies op de zere plek. Je inzichten op theoretisch en methodologisch vlak zijn erg waardevol geweest. Je hebt veel bijgedragen aan dit proefschrift. Dank je daarvoor.

Mijn collega's Jeroen Stekelenburg en Ilja Frissen verdienen ook veel dank. Allebei hebben jullie me in de eerste maanden van mijn aio-tijd wegwijs gemaakt en zorgden jullie ervoor dat ik me thuis voelde binnen de onderzoeksgroep "cognitive neuroscience". Jeroen, ik vond het erg prettig met je samen te werken en te discussiëren. Dank je voor de lol die ik met je had, voor je antwoorden op mijn vele vragen en natuurlijk voor je advies op het gebied van mode en kleding. Bij jou kon ik echt voor van alles terecht. Ilja, ik wil je bedanken voor al je hulp, voor je kritische opmerkingen en voor je enthousiasmerende betogen over alles dat met perceptie en onderzoek te maken heeft. En natuurlijk voor je gezelligheid!

Mijn twee kamergenoten, Mirjam en Ruthger wil ik beiden bedanken voor de gezelligheid en de fijne tijd. We zijn alle drie tegelijk begonnen en ook weer alledrie tegelijk en bovendien op tijd klaar, en dat is best bijzonder lijkt me. Mirjam, je was meer dan een kamergenootje. Dank je voor een hele fijne tijd, voor het plezier en zeker ook voor je luisterende oor en je steun. Ik moet zeggen dat ik het toch wel erg jammer vind dat je voorlopig niet bij TNO zult gaan solliciteren...

Eigenlijk hebben alle collega's en medewerkers binnen de afdeling psychologie ervoor gezorgd dat mijn aio-tijd vooral ook een erg leuke is tijd geworden. Eenieder die aansloot aan de lunchtafel in de kantine of aan de borreltafel in de Esplanada: het maakte het werken aan de UvT een stuk gezelliger en zeker ook interessanter door alle discussies die werden gevoerd over de meest uiteenlopende onderwerpen.

Mijn twee paranimfen, mijn lieve zus Elvira en mijn fijne vriendin Marieke wil ik bedanken erg dat jullie ervoor gezorgd hebben dat de laatste loodjes ook nog een beetje leuk zijn geworden. Dank je wel dat jullie, letterlijk en figuurlijk, achter me willen staan op een toch wel spannende dag.

Papa en mama, dank je wel voor jullie geloof in mij en voor het beste advies dat ik ooit gekregen heb: "Doe maar gewoon je best"! Dat heb ik altijd gedaan, en, misschien net als jullie, verwonder ik me er wel eens over hoever ik daarmee gekomen ben.