Hearing and seeing speech:

Perceptual adjustments in

auditory-visual speech processing

Cover design: Dennis van der Zande

ISBN: 978-90-76203-50-8

Printed and bound by Ipskamp Drukkers B.V., Nijmegen

Hearing and seeing speech:
Perceptual adjustments in
auditory-visual speech processing

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,

volgens besluit van het college van decanen

in het openbaar te verdedigen op donderdag 5 september 2013

om 13.00 uur precies

door

Patrick Henricus Eduard van der Zande

geboren op 5 februari 1986

te 's-Hertogenbosch

Promotor:

Prof. dr. A. Cutler

Copromotor:

Dr. A. Jesse (University of Massachusetts, Amherst, USA)

Manuscriptcommissie:

Prof. dr. A.H. Özyurek

Prof. dr. M.G. Swerts (Tilburg University)

Prof. L. D. Rosenblum (University of California, Riverside, USA)

The research reported in this thesis was supported by a grant from the Max-Planck-Gesellschaft zur Förderung der Wissenschaften, München, Germany.

Acknowledgements

Even though my name is the only one that appears on the front cover of this dissertation, I could obviously never have finished all of this work without the help of others. In fact, there have been many people who have contributed in some way to the work that is discussed in the coming pages. Regardless of whether their contribution was substantial or only small, I would like each and every one of these people to know that I am extremely grateful for the role that they played in the process that has now, finally, resulted in this dissertation. I cannot name everyone individually, but you are all awesome.

First and foremost, I would like to thank my day-to-day supervisor Alexandra Jesse. Alexandra, thank you for hiring me as a student assistant back in 2008. I really consider that as the start of my Ph.D. work. I was already interested in audiovisual speech perception before I started my Ph.D., as was clear from my master's thesis, but you have managed to keep me enthusiastic about the subject all these years. I am very happy that you had enough faith in me to let me do a Ph.D. with you as well. It hasn't always been easy, but I have learnt a lot during my time as a *promovendus* at the MPI and none of that would have been possible without your trust. I appreciate your positivity about our work and our results, even though at times we had no idea what to do with the data that we had collected. Being able to visit you at UMass also provided me with a great opportunity to see what life as a researcher can be like outside of the MPI.

Next, I would like to thank my promoter, Anne Cutler, for always being incredibly enthusiastic about new results coming in and for really making me believe that my findings were relevant and interesting. Anne, you've been a great motivator and I have fond memories of some of your more upbeat e-mails, recent ones containing such exclamations as "Excellent!" and "Wooheee!!" They served as a great pick-me-up when at times I felt like things weren't really going my way. Also, many

thanks to you for allowing me to do research in the stimulating, positive atmosphere that the MPI and especially the Language Comprehension group has provided over the years.

All the members of the Language Comprehension group and others who had a close connection with our group deserve a thank-you one way or another. Our biweekly group meetings were always interesting and it was inspiring to hear about everyone else's results after being so focussed on my own work. Even though I didn't present an awful lot during these meetings, I am still grateful for all the valuable comments and helpful tips that you had for me when I did. In particular, I would also like to thank Holger Mitterer for his help with using Praat and STRAIGHT. Also, Holger, whether you realise it or not, you have made my work a *lot* easier and more efficient by teaching me how to write scripts in Perl. I doubt I would've been able to finish when I did if I hadn't learnt how to make loops and ifelse statements.

I have shared an office with quite a lot of people over the years. Although there are too many to name individually, each and every one of them should know that they were an absolute joy to work with. Working as a student assistant was an incredibly fun experience and I'm extremely happy that I was allowed to take part in the student assistant parties even well after I had started my Ph.D. It was always very pleasant to be able to leave my own office and hang out in 381 for a while, just to take my mind off of things. Laurence, you were the best source of information for things going on within the institute and you took very good care of all the student assistants who came and went during your time at the MPI. I hope you are enjoying your own Ph.D. position now.

I would also like to thank Susanne Brouwer, who was one of my first officemates when I switched from being a student assistant to being a Ph.D. candidate. Susanne, you were the voice of reason when I started with my new position and still wanted to do everything by the book. Thanks for telling me that it was okay to stop struggling and to just go home when things really weren't working

out. I also want to thank my more recent officemates Katja, Jiyoun, and Wencui for creating a lively office space and for the cheerful conversation. It was always nice having someone there in the office with whom to discuss work-related and non-work-related business. Writing and drawing on our two huge whiteboards was a very welcome creative break from making stimuli and analysing data. Jiyoun, I enjoyed trying to teach you Dutch and learning Korean in return. At least now, if I ever end up in Korea, I know how to say "hello!" to people and how to ask them to lunch…

My experiments all contained auditory and visual records of various colleagues and I want to thank Jelmer, Marijt, Joost, and Matthias for serving as talking heads for my experiments. Jelmer, you are not only the person I recorded most, but you have also become a good friend. You were the first person I ever ran an experiment with at the MPI and you quickly became my go-to guy when I needed new stimuli. I'm glad you were willing to sit in the Gesture Lab for hours on end, repeating real and fictive Dutch words while staring into the camera. If it hadn't been for your inhuman skill of sitting still and not blinking, I don't think my experiments would've worked out at all. Outside of the MPI, I really enjoyed going to the movies and having BBQs together. I am also very happy that you have agreed to stand by my side during my defence. And one day I *will* beat you in squash! Marijt, you were the female voice and face in my experiments and I thank you, too, for being willing to have yourself recorded and displayed in front of my participants. You seemed to know everything about everything and you have always been really helpful, even after confessing that you sometimes hated the fact that everyone always just asked you when they needed to know something. You deserve additional mention for your work as Ph.D. representative (a job that really suited you well!), which ultimately resulted not only in the beautiful new design of our dissertations but also in an actual pay raise for Ph.D. students! You, too, will be at my side during my defence and I'm extremely grateful for that.

I often needed equipment for recording my stimuli and I want to thank the Technical Group for their help in getting me the things that I needed. I was generally able to pick up everything faster than I had hoped, which really helped me keep up a nice pace with recording stimuli and running experiments. Despite the fact that my setups weren't always technically sound (using extension cord after extension cord, for instance), I'm happy that you guys were always OK with providing me with whatever I requested. Particular thanks go to Alex Dukers for allowing me to send him e-mails directly, rather than having to go through the official channels.

Dan wil ook nog graag een aantal mensen bedanken in het Nederlands. Ten eerste wil ik Reinier bedanken, omdat ik erg blij ben dat ik in hem een concertmaatje heb gevonden. Het is enorm fijn dat je altijd bent te porren om naar een of ander herrieconcert te gaan in De Onderbroek of Willemeen, of waar dan ook. Als je geld hebt tenminste… Je hebt me geestelijk gezond gehouden door me mee naar buiten te nemen voor avonden gevuld met harde muziek en goedkoop bier. Daar ben ik je dan ook enorm dankbaar voor. Ook is 't altijd gezellig als je komt eten en ben ik blij dat ik met je kan praten over slechte televisieseries.

Mijn thuisfront mag natuurlijk ook niet ontbreken en ook hen ben ik veel dank verschuldigd. Dennis, bedankt voor 't onderwerpen van de prachtige kaft die mijn proefschrift compleet heeft gemaakt. Mama, papa, bedankt voor jullie steun en vertrouwen. Ik weet dat ik niet altijd even goed kon uitleggen wat er gaande was en hoe alles in elkaar stak op mijn werk, maar zoals jullie zien is alles uiteindelijk toch goed gekomen. Bedankt voor alle warmte en het altijd prettige gevoel van thuiskomen in de weekenden.

En als allerlaatste wil ik graag Juul nog bedanken. Bedankt voor je zorgzaamheid en je goede inzichten, maar vooral bedankt, omdat 't leven een stuk leuker is sinds ik het met jou deel.

Table of contents

Chapter 1:

Introduction

1. Introduction

Over the course of our entire lives, we communicate with a large variety of people of different ages and different backgrounds. Every person we talk to tends to produce sounds in a uniquely personalised manner, often referred to as their idiolect. The speaker's physical features, their geographical background, and their socio-economic upbringing all shape this personal style of speaking (Fant, 1973; Foulkes & Docherty, 2006; Ladefoged, 1980; Laver & Trudgill, 1979; Peterson & Barney, 1952). Speakers vary not only in the way they produce sounds but also in other aspects, such as their speaking rate and in their choice of words (Miller, Grosjean, & Lomanto, 1984). Imagine that a person's idiolect is the spoken equivalent of their handwriting: while everyone's handwriting shows subtle and not-so-subtle differences, certain conventions on how the characters are to be shaped will still be adhered to. As readers, we are typically able to see the intended message despite the variety in the details of its surface form. Similarly, as listeners, we are highly capable of dealing with variations in speech and have different mechanisms available that help us extract the intended message despite the idiosyncrasies that occur within the signal.

The studies discussed in this thesis provide new insights on how listeners adjust their perceptual system to speakers' idiosyncrasies with the goal of streamlining the interpretation of speech. The main focus of this thesis is adjustment to speakers on the basis of audiovisual speech input. Face-to-face conversations make up a large portion of our everyday interaction with others. Listeners have repeatedly been shown to be able to benefit from combined auditory and visual speech input. Despite the use of visual-only and audiovisual speech materials, the terms "listener" and "perceiver" are used interchangeably throughout this thesis to refer to the person who is interpreting speech, regardless of the modality in which the materials are presented. This introductory chapter provides a discussion of the previous work on audiovisual speech perception and speaker familiarity. The chapter ends with a short overview of the experiments conducted in this thesis.

1.1. Audiovisual speech perception

As a speaker produces speech, listeners process the incoming speech signal to unravel what is being said. Often, auditory speech provides enough information for the interpretation of a speaker's utterance. In face-to-face conversations, however, listeners also process the speech information they can obtain from the speaker's talking face. The information provided by the visual speech helps to improve the perception of speech. Combined auditory-visual speech is therefore always more informative than auditory speech alone. The benefit of audiovisual speech perception over auditory-only speech perception is not limited to listeners who have difficulty hearing (Grant, Walden, & Seitz, 1998; Kaiser, Kirk, Lachs, & Pisoni, 2003): all listeners show improvements in recognition when presented with audiovisual speech regardless of their hearing acuity or their age (Jesse & Janse, 2009; Jesse, Vrignaud, Cohen, & Massaro, 2000/2001; MacDonald & McGurk, 1978; Macleod & Summerfield, 1987).

The usefulness of combined auditory and visual speech signals is particularly clear in situations where the auditory speech signal is degraded (Sumby & Pollack, 1954). When it is particularly difficult to understand a speaker from listening alone due to the background noise (Macleod & Summerfield, 1987), visual speech can facilitate the detection of auditory speech in noise (Bernstein, Auer, & Takayanagi, 2004; Grant & Seitz, 2000) and can be particularly useful for understanding the speaker (Middelweerd & Plomp, 1987; Summerfield, 1992). Being able, in such cases, to see as well as hear the speaker talk will result in more information being available for speech perception and will therefore result in improved recognition of speech. Visual speech provides information about the phonetic segments that occur in an utterance but also contains suprasegmental information about prosody (Krahmer, Ruttkay, Swerts, & Wesselink, 2002; Krahmer & Swerts, 2004; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Visual speech information is not, however, only used in cases where auditory speech is difficult to understand. In fact, information from the two speech modalities

is integrated automatically whenever both are available (Massaro, 1987, 1998). Such integration occurs even when listeners are explicitly told to focus on one of the two signals, indicating that this process of integration is not under conscious control (McGurk & MacDonald, 1976; Reisberg, McLean, & Goldfield, 1987; Soto-Faraco, Navarra, & Alsius, 2004).

Speakers produce different sounds by changing the positioning of their articulators and being able to see these movements in the speaker's face informs the listeners about the sounds that a speaker likely produced. It is this inherent link between auditory speech and visual speech that makes the combined auditory-visual speech input such a strong source of information for speech perception (Yehia, Rubin, & Vatikiotis-Bateson, 1998). Auditory speech and visual speech are integrated because of this common, shared source, if they are linked in time and space. Despite this shared origin, auditory speech and visual speech are not equally intelligible when presented separately. The number of phonemes that can be distinguished visually, for instance, is smaller than the number of phonemes that can be distinguished auditorily. Fewer visual phonetic categories are recognised than auditory phonetic categories (Owens & Blazek, 1985; Van Son, Huiskamp, Bosman, & Smoorenburg, 1994; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). These visual phonetic categories consisting of phonemes that are particularly difficult to distinguish visually are called visemes (Fisher, 1968). The information that is available in auditory speech and visual speech can be redundant and complementary (Grant et al., 1998; Jesse & Massaro, 2010; Walden, Prosek, & Worthington, 1974). When the information in the two signals is redundant and equally likely to be perceived from either input source, this provides additional strength to the interpretation of the utterance. But the information in the auditory signal and the visual signal can also be complementary because certain cues for sounds are more easily distinguished in one modality than in the other. Auditory speech contains strong cues for voicing and the manner of articulation of a phoneme (e.g., frication), which are difficult to detect visually due to the fact that they occur in an internal part

of the vocal tract. One exception here is lip rounding, which involves the speaker's mouth and thus is visually salient (Breeuwer & Plomp, 1986). Visual speech, on the other hand, contains clear cues for the place of articulation (e.g., bilabial) due to the visible movements of the articulators and this distinction is more difficult to make in auditory speech (Breeuwer & Plomp, 1986). To produce a voiceless bilabial plosive /p/, for instance, speakers close their lips and build up air pressure behind this closure, which is subsequently released. A sustained closure of the lips results in an auditory silence that may not be very informative about the place of articulation. The visible movements and closure in the speaker's face provide clear evidence for the place of articulation, however. Articulatory cues are also available earlier in visual speech than in auditory speech, with movements of the mouth often preceding the occurrence of sound (Jesse, 2005; Jesse & Massaro, 2010).

Our communication with others consists largely of face-to-face interactions in which we are able to both hear and see the person with whom we are speaking. Listeners cope with the problems that may arise in the perception of speech, for instance due to variations in the signal, and in order to find out more about how such problems can be overcome it is imperative to further our knowledge of how listeners perceive speech that is presented audiovisually. Audiovisual speech provides the most complete source of speech information with which listeners can be presented and therefore also provides a large amount of information about the speaker to which a listener can adjust.

1.2. Speaker familiarity

Although the substantial variation that occurs within natural speech could make speech signals ambiguous, we know that listeners are quite capable of understanding the people with whom they interact. Listeners are able to extract the intended message from an utterance even when produced by a speaker whom they have not previously encountered. The auditory speech input provides listeners with information about the speaker's idiolect and the variation that occurs in the auditory

signal may actually be beneficial to the listener as well as being problematic (Pisoni, 1993). Listeners are able to learn from exposure to a speaker's speech and words that are repeated by the same speaker are recognised faster and more accurately, for instance, than words repeated in a different voice (Palmeri, Goldinger, & Pisoni, 1993). Listeners thus store information about the voice of a speaker in long-term memory and can use this knowledge on subsequent encounters, which facilitates the perception of speech produced by familiar speakers (Nygaard, Sommers, & Pisoni, 1994).

Although having information about a speaker's idiolect stored in long-term memory is useful, the encoding of this information requires additional resources. These resources will generally be drawn away from other processes occurring simultaneously. The increased demand on cognitive resources when listeners process unfamiliar voices results in performance being worse when hearing different speakers in succession than when hearing the same speaker throughout (Martin, Mullennix, Pisoni, & Summers, 1989; Mullennix, Pisoni, & Martin, 1989; Palmeri et al., 1993). In return, however, processing is faster and more accurate for speakers to whom listeners have been familiarised once the speaker-specific information has been stored. The benefit for perception of words produced by a single speaker stems from the fact that voice information does not have to be encoded for every utterance of a familiarised speaker and this information is used to adjust the analysis of the incoming speech signal.

The encoding of speaker-specific information to long-term memory occurs automatically. Details about the speaker's idiolect are stored regardless of the task that listeners perform during their exposure to the speaker's voice and even in the absence of explicit instructions. Speech perception shows benefits from familiarity with a speaker's voice, even when the initial exposure task did not involve the identification of the speech (Nygaard & Pisoni, 1998). For example, when listeners were taught the names of novel voices they heard during exposure, their subsequent recognition of speech from these newly familiarised speakers during test showed

improvements (Nygaard et al., 1994). Mere exposure to a speaker's voice is therefore sufficient for information about the idiosyncrasies to be stored in memory, even without explicit instructions to do so. Furthermore, explicit instructions for listeners to focus on the identity of the speaker does not improve recognition (Palmeri et al., 1993), again indicating that the encoding of speaker information is not modulated by specific task demands.

Familiarity with a speaker's voice affects the processing of all subsequent speech produced by the same speaker, rather than only facilitating the perception of previously perceived words (Nygaard et al., 1994; Pisoni, 1993). This generalisation of speaker familiarity suggests that listeners acquire details about how speakers produce particular sounds, improving identification of all words that contain such sounds. The information that is stored in long-term memory shows some specificity for the exposure context, however. When listeners are exposed to a speaker's idiosyncrasies through sentence-length material, the knowledge they acquire does not generalise very well to the identification of words in isolation (Nygaard & Pisoni, 1998). Therefore, listeners may acquire information about a speaker's voice that is dependent on the specific context in which it was presented and the stored information appears to not generalise readily to other contexts.

Similar effects of speaker familiarity have been observed when visual speech information is involved. Words are faster and more accurately identified from visual-only speech, that is, when speakers are only seen and not heard, when the same speaker is presented than when the speaker is different from trial to trial (Yakel, Rosenblum, & Fortier, 2000). Task demand and cognitive processing loads are thus higher in visual-only speech with multiple speakers, suggesting that listeners are sensitive to variations in visual speech as well as in auditory speech (Sheffert & Fowler, 1995). Listeners are not only able to cope with this variation but can use it to benefit the future recognition of speech by the same speaker. Adjustments after familiarisation to a speaker's idiolect can generalise across modalities: Exposure to a speaker's idiolect through visual-only speech identification can improve subsequent

recognition of the same speaker's production of auditory speech (Rosenblum, Miller, & Sanchez, 2007). This finding indicates that speaker-specific information may be general enough to transfer across modalities so that even information from one modality is sufficient to adjust our expectations of a speaker's sounds in another modality (Rosenblum, 2008). Information about previously perceived speakers results in changes in the perceptual system that can ultimately facilitate processing of speech regardless of the modality in which the speech is presented. Therefore, while it may initially slow down processing, being able to learn about idiolects is a very useful tool for speech perception.

1.3. Perceptual learning

An important line of research on adjustments to speakers in the last ten years has focused on perceptual learning or phonetic retuning (Bertelson, Vroomen, & De Gelder, 2003; Norris, McQueen, & Cutler, 2003). Perceptual learning studies have been used to demonstrate that listeners adjust their phonetic categories on the basis of the speaker-specific information with which they are presented. When the input contains a phoneme that can be interpreted as belonging to two separate categories, the boundary between those categories can be adjusted so that the previously ambiguous phoneme can be assigned to the correct category. Listeners thus dynamically change the boundaries between their phonetic categories in order to facilitate the identification of speech. Changes in category boundaries are only made when the listener is presented with consistent information about the direction in which the boundaries need to be shifted.

In auditory-only speech perception, the boundary between two auditory phonetic categories can shift on the basis of lexical knowledge, for instance (Norris et al., 2003). Due to the speaker-specific variations in speech, a particular sound may be difficult to interpret in a speaker's idiolect. Such idiosyncratic sounds can generally be disambiguated by the lexical context in which they are presented, however, and such knowledge about which words occur in the language provides sufficient

information for correct identification. When the word-final fricative /s/ in the word *platypus* is replaced by an ambiguous fricative between /f/ and /s/, for instance, the lexical context of the word can inform the listeners the odd sound was actually the speaker's idiosyncratic realisation of the phoneme /s/. Hearing the same ambiguous fricative in the context of *giraffe*, however, listeners' lexical knowledge will make them identify the idiosyncratic sound as being an /f/. Continued exposure to the speaker-specific idiosyncrasy in a consistent lexical context will result in listeners retuning the boundary between their categories for /f/ and /s/ (Norris et al., 2003). Listeners subsequently identify the idiosyncratic sound as belonging to the /s/ category when they were exposed to the sound in the context of *platypus*, or as belonging to the /f/ category after exposure to *giraffe*, even when presented outside of its original lexical context. This does not change the way in which listeners perceive the idiosyncrasy, just the category to which they learn to assign the sound.

The retuning of category boundaries occurs so that the idiosyncratic sounds can be assigned to the proper phonetic category and these shifts facilitate the subsequent identification of speech from the same speaker. The effect of phonetic retuning is different from the adjustments that occur in the perceptual system as a result of repeated exposure to unambiguous speech. When listeners are repeatedly presented with the same unambiguous speech material, this will result in them identifying fewer ambiguous sounds as being part of the exposure category. Adaptation caused by unambiguous speech, also called selective adaptation, may be due to overexposure to a particular sound rather than an effort to facilitate the processing of problematic speech (Diehl, 1975; Eimas & Corbit, 1973; Roberts & Summerfield, 1981; Samuel, 1986).

Like lexical knowledge, visual speech information can also cause shifts in the boundaries between auditory phonetic categories (Bertelson et al., 2003). Auditory idiosyncrasies can be disambiguated when presented together with unambiguous visual speech information. An utterance containing an ambiguous plosive between /b/ and /d/, combined with a full visual closure of the speaker's lips, can be readily

interpreted as having been intended as a /b/. The visible movements of the articulators provide contextual information by restricting the possible interpretations for the idiosyncrasy. Disambiguation by visual speech input also results in the retuning of the boundaries between auditory categories for /b/ and /d/ (Bertelson et al., 2003) and adjustments to the boundary again occur so that the idiosyncratic sound can be assigned to the correct category given the disambiguating information. Both lexically guided and visually guided retuning result in comparable changes to the auditory phonetic category boundaries (Van Linden & Vroomen, 2007).

Visual speech itself also displays idiosyncrasies that may lead the visual speech signal to be ambiguous. In the case of visual-only idiosyncrasies, listeners can use unambiguous auditory speech to disambiguate the visual idiosyncrasies and to retune the visual phonetic categories (Baart & Vroomen, 2010). Accordingly, both auditory phonetic categories and visual phonetic categories can be shifted given sufficient disambiguating information and hence facilitate the processing of speech by making sure that the idiosyncratic sound falls within the correct phonetic category. In other words, ambiguities in one modality can be resolved by unambiguous speech in another modality and the boundaries between phonetic categories are adjusted in whichever modality the ambiguity occurred.

Shifts in phonetic category boundaries occur after a relatively short exposure (Kraljic & Samuel, 2007) and the retuning affects all subsequent processing for speech from the same speaker containing that particular phoneme (Jesse & McQueen, 2011; McQueen, Cutler, & Norris, 2006). In certain cases, the retuning of phonetic boundaries can even affect the perception of speech that was produced by a different speaker, although here it is the nature of the ambiguous phoneme that determines whether speaker-specific knowledge can be generalised (Eisner & McQueen, 2005; Kraljic & Samuel, 2006). Phonemes that sound similar across speakers and thus contain little speaker-specific information (i.e., plosives) allow for generalisation, while phonemes that sound very different across speakers and thus contain a large amount of speaker-specific information do not (i.e., fricatives). In the case of plosives,

where generalisation across speakers is possible, adjustments made for one speaker can facilitate the processing of speech from other speakers (Kraljic & Samuel, 2006). The effects of phonetic retuning are detected even after an intervening period, indicating that they do not dissipate quickly after exposure (Eisner & McQueen, 2006; Vroomen & Baart, 2009; Vroomen, Van Linden, Keetels, De Gelder, & Bertelson, 2004). Depending on whether the auditory categories or the visual categories were retuned, the effects are still visible up to 24 hours after the initial familiarisation.

Once a category boundary has been shifted, exposure to another speaker does not immediately reset it (Kraljic & Samuel, 2005). This may be partially due to the fact that the retuning are a specific shift forced by the information in the exposure, not simply a weakening of the boundaries (Maye, Aslin, & Tanenhaus, 2008). Perceptual boundaries are reset when the same speaker is heard producing canonical, non-idiosyncratic realisations of the previously idiosyncratic phoneme for which the boundaries were adjusted (Kraljic & Samuel, 2005). In fact, when the speaker is first heard producing canonical forms and later produces idiosyncratic realisations of the same sound, phonetic retuning does not occur. Similarly, when listeners can see that the idiosyncratic sounds are due to an outside source (for instance, a pen in the speaker's mouth), no adjustments are made in the perceptual system (Kraljic, Samuel, & Brennan, 2008). In these cases, listeners are able to detect that an idiosyncratic utterance were due to the external factors and are not actually part of the speaker's idiolect (Kraljic & Samuel, 2005; Kraljic et al., 2008).

The retuning of phonetic categories is clearly a strong example of how listeners are able to adjust to the variations that occur in natural speech. Adjustments are made quickly and facilitate subsequent processing of speech from a familiar speaker, regardless of the specific words that this speaker produces. This lasting effect of retuning is robust and resilient but can be prevented or undone if the listener is presented with outside explanations for the occurrence of the idiosyncratic sound.

1.4. Outline of the thesis

The goal of this thesis is to provide new insights into how listeners are able to deal with speaker-specific variations that occur in audiovisual speech perception. How do listeners adjust to speaker-specific idiosyncrasies that occur in audiovisual speech and do changes that occur in the perceptual system as a result of exposure to a particular speaker also affect the perception of speech produced by another speaker? Furthermore, what can these specific adjustments tell us about the nature of the information that is stored in memory? Finally, it was also investigated whether previously obtained knowledge about a speaker's idiolect improves implicit and explicit memory for repeated words.

In Chapter 2, the focus is on the phonetic retuning of visual category boundaries guided by lexical information. Previous research has indicated that lexical information can guide the retuning of auditory category boundaries (Norris et al., 2003) and that auditory speech information can result in changes in visual phonetic categories (Baart & Vroomen, 2010). The study reported in Chapter 2 establishes whether lexical information can guide the retuning of visual phonetic categories. First, in Experiment 2.1, the retuning of visual phonetic categories is tested using audiovisual materials that are ambiguous in both the auditory and the visual speech modality. The results of Experiment 2.2 are used to discuss whether the retuning of visual phonetic category boundaries can be indirectly due to changes in auditory phonetic categories. In other words, can a shift in the auditory category boundary later result in the auditory speech signal serving as a disambiguating cue for the visual idiosyncrasy?

In the study reported in Chapter 3, phonetic retuning (Experiment 3.1) and selective adaptation (Experiment 3.2) are examined. These effects both occur as a result of previously perceived speech and lead to two very different changes in the perceptual system. Listeners were exposed to audiovisual speech that either consisted of an ambiguous auditory signal combined with an unambiguous visual signal (Experiment 3.1) or to audiovisual speech that was fully unambiguous

(Experiment 3.2). Listeners in both experiments were subsequently tested on auditory-only speech that was produced by either the exposure speaker or by a novel speaker. The two experiments in Chapter 3 investigate whether the effects of phonetic retuning and selective adaptation influence only the subsequent perception of speech produced by the familiarised speaker or whether they also influence the perception of speech from a different speaker than the one to whom listeners were initially exposed.

In Chapter 4, cross-modal(ity) priming is used in order to investigate whether exposure to a speaker through auditory speech can subsequently facilitate the perception of visual-only speech produced by that same speaker. The two experiments reported in Chapter 4 specifically focus on the effects of word repetition and speaker repetition on the identification of words in a long-term priming paradigm. In Experiment 4.2, an additional recognition memory task is used to determine whether word repetition and speaker repetition affect explicit memory of the repeated items.

Finally, Chapter 5 provides a short summary and overview of the major findings reported in the experimental chapters.

Chapter 2:

Lexically guided retuning of visual phonetic categories

Van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *Journal of the Acoustical Society of America, 134*(1), 562-571.

Abstract

Listeners retune the boundaries between phonetic categories to adjust to individual speakers' productions. Lexical information, for example, indicates what an unusual sound is supposed to be, and boundary retuning then enables the speaker's sound to be included in the appropriate auditory phonetic category. In this study, it was investigated whether lexical knowledge that is known to guide the retuning of auditory phonetic categories, can also retune visual phonetic categories. In Experiment 1, exposure to a visual idiosyncrasy in ambiguous audiovisually presented target words in a lexical decision task indeed resulted in retuning of the visual category boundary based on the disambiguating lexical context. In Experiment 2 it was tested whether lexical information retunes visual categories directly, or indirectly through the generalisation from retuned auditory phonetic categories. Here, participants were exposed to auditory-only versions of the same ambiguous target words as in Experiment 1. Auditory phonetic categories were retuned by lexical knowledge, but no shifts were observed for the visual phonetic categories. Lexical knowledge can therefore guide retuning of visual phonetic categories, but lexically guided retuning of auditory phonetic categories is not generalised to visual categories. Rather, listeners adjust auditory and visual phonetic categories to talker idiosyncrasies separately.

1. Introduction

In everyday communication, listeners encounter a variety of talkers, and all of them may pronounce the sounds of their native language in their own specific, idiosyncratic way. Such variation between speakers can arise from physiological differences (Laver & Trudgill, 1979), or because speakers have different dialectal and sociological backgrounds (Foulkes & Docherty, 2006). Given proper disambiguating information, however, listeners quickly and effectively adjust phonetic category boundaries to incorporate speakers' idiosyncratic realisations of sounds into the correct phonetic categories (Baart & Vroomen, 2010; Bertelson, Vroomen, & De Gelder, 2003; Jesse & McQueen, 2011; Norris, McQueen, & Cutler, 2003). In face-to-face communication, listeners also make use of visual information about their interlocutors' articulation, and in doing so they draw on visually defined categories for individual phonemes (Massaro, 1998; Van Son, Huiskamp, Bosman, & Smoorenburg, 1994). Idiosyncratic articulations may also require the retuning of these visual phonetic categories. Simultaneously presented auditory information that disambiguates the sound can guide such retuning (Baart & Vroomen, 2010). Suppose, however, that an idiosyncratic articulation results in a sound being simultaneously both visually and auditorily ambiguous. In that case, the listener may still use lexical knowledge to guide retuning. But is one retuning operation then needed, or two? We investigate here whether lexical knowledge (known at least to retune auditory category boundaries: Norris et al., 2003) can lead to a retuning of visual phonetic categories in the absence of explicit auditory disambiguation. We further test whether retuning of visual phonetic categories can occur through generalisation across modalities. Can retuning of auditory phonetic categories on the basis of lexical information also result in shifts of visual category boundaries?

Norris and colleagues (2003) showed that knowledge about the words of listeners' native language not only disambiguates idiosyncratic sounds but also results in shifts in listeners' auditory phonetic category boundaries. Dutch listeners were presented with either /s/-final words such as *radijs* "radish", or /f/-final

words such as *olijf* "olive" where the final fricative sound was replaced with an ambiguous sound between /s/ and /f/. Despite this alteration, listeners accepted these words in lexical decision. In a subsequent categorisation task, listeners who had been exposed to the ambiguous sound in words normally ending in /s/ categorised more sounds from an /s/-/f/ continuum as /s/ than listeners exposed to the same sound in words normally ending in /f/. Thus reference to existing knowledge allows category boundaries to be rapidly adjusted to incorporate an ambiguous sound into the appropriate phonetic category. This lexically guided retuning can be speaker-specific (Eisner & McQueen, 2005), and is stable in that its effects last at least for 24 hours (Eisner & McQueen, 2006). Besides for fricatives, as in these studies, this retuning has been demonstrated for stop consonants (Kraljic & Samuel, 2006) and liquids (Scharenborg, Mitterer, & McQueen, 2011), as well as for lexical tone in Mandarin (Mitterer, Chen, & Zhou, 2011).

Importantly, retuning facilitates speech recognition in any situation where a similar idiosyncrasy is encountered. The effect of lexically guided retuning for auditory phonetic categories generalises across word-internal positions and also generalises to novel words (Jesse & McQueen, 2011; McQueen, Cutler, & Norris, 2006; Mitterer et al., 2011; Sjerps & McQueen, 2010). Listeners who were exposed to an ambiguous fricative between /f/ and /s/ in word-final position showed, for example, boundary shifts in line with their exposure even when the ambiguous fricative occurred in word-initial position (Jesse & McQueen, 2011). In another study, listeners performed a cross-modal priming task at test that included auditory primes ending in the ambiguous fricative. The ambiguous auditory primes, e.g., /naɪ?/, could be interpreted as either an /f/-final word ("knife") or an /s/-final word ("nice"). The pattern of priming from these ambiguous auditory tokens revealed that they were interpreted by listeners in line with the listeners' prior exposure (McQueen, Cutler et al., 2006; Sjerps & McQueen, 2010). Phonetic retuning thus

allows listeners to deal with the considerable variability that speakers show in their pronunciation of the sounds of their native language.

Communication is not a purely auditory phenomenon, however, and spoken interaction also provides visual information, for instance concerning articulatory movements. In face-to-face communication, listeners automatically combine information obtained from hearing and seeing a speaker (Massaro, 1987, 1998). Visual speech affects identification even when listeners are instructed to disregard talkers' mouth movements (Massaro & Cohen, 1983; McGurk & MacDonald, 1976). This use of visual speech information is typically beneficial to the listener, as it improves the intelligibility of a speaker significantly (e.g., Helfer & Freyman, 2005; Jesse, Vrignaud, Cohen, & Massaro, 2000/2001; Macleod & Summerfield, 1987; Reisberg, McLean, & Goldfield, 1987; Spehar, Tye-Murray, & Sommers, 2008). Bimodal speech perception is especially useful when the input in one modality is difficult to interpret (Sumby & Pollack, 1954). The information provided by the two modalities is redundant but also complementary in that phonetic features that are difficult to distinguish in one modality are often more easily distinguished in the other modality (Grant, Walden, & Seitz, 1998; Jesse & Massaro, 2010; Summerfield, 1987; Walden, Prosek, & Worthington, 1974). Because of this, audiovisual speech recognition performance often exceeds the simple addition of auditory-only and visual-only performances (Massaro & Cohen, 1983; Massaro & Friedman, 1990). The benefit of bimodal speech perception over unimodal perception decreases, for example, with increased redundancy between the information from the two modalities (Grant et al., 1998).

The influence of visual speech input goes beyond simple facilitation of recognition through disambiguation. Like lexical information, visual speech input guides the retuning of auditory phonetic categories (Bertelson et al., 2003). Simultaneously presented visual speech can disambiguate an acoustically ambiguous plosive between /b/ and /d/ by indicating whether the presented sound was a bilabial or an alveolar sound. Listeners who have been exposed to audiovisual

stimuli containing an auditory idiosyncrasy show boundary shifts that are in line with the visual disambiguating information in a subsequent auditory-only categorisation task. Auditory phonetic categories are thus retuned both by lexical information and by simultaneously presented visual speech information, the effects of which have also been shown to be statistically similar in size (Van Linden & Vroomen, 2007).

Visual speech itself can also be idiosyncratic, however. Familiarity with the visual speech of a talker can improve subsequent recognition of the talker's visual and auditory speech (Rosenblum, Miller, & Sanchez, 2007; Rosenblum, Yakel, & Green, 2000; Yakel, Rosenblum, & Fortier, 2000). Participants recognised visual speech better, for example, when the same speaker was presented throughout a visual-only recognition task than when multiple speakers were shown (Yakel et al., 2000). Listeners can also match a speaker's face producing a sentence to their subsequently presented voice, even when the linguistic content of the visual and auditory speech differ (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007). These results suggest that listeners adjust to the visual idiosyncrasies of a speaker. Auditory speech information can guide the adjustment to visual idiosyncrasies, when these make visual productions of sounds ambiguous. Baart & Vroomen (2010) presented listeners with videos of a talker producing /o?so/, where /?/ was a visually ambiguous nasal between /m/ and /n/. Audiovisual stimuli were created by combining the ambiguous visual speech input with natural auditory /omso/ or /onso/ tokens. Exposure to these audiovisual stimuli resulted in retuning of the visual phonetic categories. Auditory information thus guides retuning of visual phonetic categories, confirming that speech information from one modality can change category boundaries in the other modality.

However, listeners may also apply lexical knowledge to adjust visual phonetic categories, either by using lexical knowledge to retune visual categories directly, or by applying what they learn about a talker's auditory speech to adjust

their expectations about the talker's visual speech. Applying lexical information to audiovisual speech could well be useful for listeners, as idiosyncrasies do not necessarily occur only in one modality at a time. In fact, given the links between visible articulatory movements and the resulting auditory sounds (Yehia, Rubin, & Vatikiotis-Bateson, 1998), idiosyncrasies that are both auditorily and visually expressed are probable. In such cases, with both modalities containing an idiosyncrasy, there would be no opportunity for one modality to guide retuning of phonetic categories in the other. In Experiment 1, we tested whether lexical knowledge can disambiguate audiovisually idiosyncratic speech and whether visual phonetic categories can be retuned on the basis of this lexical knowledge.

We also tested whether the retuning of visual phonetic categories can occur through generalisation across the modalities. If auditory and visual phonetic categories are tightly linked, then listeners should be able to retune their visual categories even if no visual information about the idiosyncrasy was present during exposure. The retuning of auditory phonetic categories would generalise across modalities and therefore indirectly affect visual phonetic categories. Visual-only exposure to the speech of a particular speaker has been shown to facilitate subsequent recognition of that speaker's auditory-only speech, both in a long-term priming task and in a sentence-recognition task (Kim, Davis, & Krins, 2004; Rosenblum et al., 2007). Rosenblum and colleagues (2007), for instance, asked listeners to lip-read a speaker for about one hour before being asked to recognise speech in noise. Listeners who heard the same speaker in the recognition task as they had seen during the exposure task performed better than listeners who heard a different speaker in the two tasks. Listeners are thus able to extract speaker-specific information from one modality and apply it to the recognition of speech in another modality. Transfer of speaker-specific knowledge across modalities has not yet been shown for phonetic retuning, however, and it remains unclear whether changes in the auditory phonetic categories could also bring about changes in the visual phonetic categories. (Certainly unambiguous auditory information can guide the

retuning of visual categories; Baart & Vroomen, 2010). In Experiment 2, we therefore tested the possibility for lexically guided retuning of auditory phonetic categories to generalise across modalities. Visual category boundaries would then be affected by lexical information, even though the listener had not received visual information about the speaker's idiosyncrasy.

Thus in Experiment 1, two groups completed multiple repetitions of an audiovisual lexical decision task, each directly followed by visual-only categorisation. During the lexical decision task, one group heard and saw an ambiguous speech token between /p/ and /t/ that replaced all word-final /p/ tokens. Another group heard and saw the same ambiguous token replacing natural /t/ tokens. In a subsequent categorisation task, both groups categorised steps from a visual-only Dutch nonword continuum from /so:p/ to /so:t/. In Experiment 2, exposure was as in Experiment 1, but both groups only heard the exposure speaker. In the categorisation test phases, both groups again categorised steps from the visual /p/-/t/ continuum. At the end of Experiment 2, both groups then also categorised steps from an auditory /p/-/t/ continuum. If lexical knowledge (directly or indirectly) retunes visual phonetic categories, then we should observe a shift in the visual phonetic boundaries in Experiment 1. If lexically guided retuning of auditory phonetic categories further generalises across modalities, a similar shift should be seen in Experiment 2, despite the absence of visual speech information during the lexical decision task. This would mean that lexical knowledge retuned auditory categories, which in turn changed the visual categories.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Forty-two native speakers of Dutch (average age 20.5 years; six males) were paid for their participation. All participants reported normal hearing and had normal

or corrected-to-normal vision. Two participants were excluded due to their insensitivity to the auditory-only continuum in the pretest. Another 10 participants (four in the /p/-exposure group and six in the /t/-exposure group) were excluded for failing to exceed a threshold of 50 percent correct 'word' responses to the ambiguous target words on the lexical decision task. The final data set that was analysed consisted of data from 30 participants, from 16 in the /p/-exposure group and from 14 in the /t/-exposure group. Fifteen additional participants from the same population took part in a visual-only pilot experiment.

2.1.2. Materials

Four /p/-final (*hoop*, *kroop*, *zoop*, and *siroop*) and four /t/-final Dutch words (*groot*, *schoot*, *schroot*, and *vergroot*) were selected as target words for the exposure phase. None of these eight target words formed a word when its coda was replaced with any other phoneme from the same viseme category (Van Son et al., 1994; e.g., *hoop* is a Dutch word, but *hoot*, *hoob*, and *hoom* are not) or with the respective other plosive. Target words contained no other phonemes from the relevant viseme categories and no other instances of /p/ or /t/. In both word sets, one target word was disyllabic and the other three were monosyllabic. Word sets were matched on their mean frequency, number of syllables, and on their lexical stress patterns using the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993). Eight phonotactically legal nonsense words were created that ended in either /f/ or /x/. These eight nonsense words contained no phonemes from the viseme categories of the target plosives. In all 16 items (eight target words and eight nonsense words) the same vowel, /o:/, preceded the final phoneme. For the categorisation tasks, the nonsense words /so:p/ and /so:t/ were used.

A male native speaker of Dutch was video recorded with a Sony DCR-HC1000E camera. Audio was recorded with two standalone Sennheiser microphones. Videos showed the speaker's head and the top of his shoulders. The

speaker produced the target words both with their natural word-final plosive and with the alternative plosive (e.g., the Dutch word *kroop* and its nonsense word counterpart *kroot*). The same speaker also produced the eight nonsense words for the lexical decision task and the *soop* and *soot* items for the categorisation tasks. All items were recorded in pairs and the talker was instructed to avoid list intonation. Videos were digitised as uncompressed 720 × 576 .avi (audio video interleave) files in PAL format. Audio sampling rate was 44.1 kHz.

We created an auditory-only continuum and a visual-only continuum using the same audiovisual *soop* and *soot* tokens for both continua. The visual-only continuum was created for the visual-only pretest and posttests. The auditory-only continuum was presented in the auditory-only pretest that was conducted to find each individual participant's most ambiguous auditory step ($A_?$). The selected sound $A_?$ appeared in all ambiguous target words for that participant during exposure. It was presented together with a visually ambiguous final plosive $V_?$ in these words. The ambiguous visual token was the same across participants but different for each target word.

a. Auditory-only pretest materials

An audiovisual token of each of *soop* and *soot* was selected based on how well the two tokens could be merged visually without causing any noticeable blurring of the speaker's facial features and facial contour. The auditory signal from both tokens was extracted and edited using Praat (Boersma & Weenink, 2006). The word-final plosives were excised by removing all sound up to the first zero crossing of the release burst. The releases of the two plosives were then morphed using the STRAIGHT signal-processing package (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999) for Matlab (The MathWorks, Inc.). This resulted in 21 individual plosive releases changing in equal 5% steps from an unambiguous auditory /t/ release (0% /p/) to an unambiguous auditory /p/ release (100% /p/). In order to provide an unbiased context for the edited releases, an ambiguous *soo* token was

created by removing the closure duration and the release from the auditory *soop* and *soot* tokens. The two resulting *soo* tokens were then morphed in a 7-step continuum with STRAIGHT. The middle step (step 4) was selected as the ambiguous context and was then combined with all 21 morphed releases. Since neither the ambiguous context nor the morphed releases contained a closure duration, a stretch of complete silence was added to these continuum steps in Praat. This artificial closure duration was manipulated to be the same duration as the average duration of the closure for /p/ and /t/ in the original *soop* and *soot* tokens (1652 ms and 1542 ms, respectively; 1588 ms for the continuum steps).

b. Visual-only pretest and posttest materials

The audiovisual tokens that were used to create the visual-only *soop-soot* continuum were the same as for the auditory-only continuum. To create the visual-only continuum, the video tracks of the *soop* and *soot* tokens were edited using Adobe Premiere CS3. These video tracks were overlaid and the opacity level of the /p/ video was systematically varied. A clip with 0% opacity for the /p/-final video shows the speaker producing an unambiguous /t/, while a clip with 100% opacity for the /p/-final videos shows the speaker producing an unambiguous /p/. A 21-step visual-only continuum was created that ranged from 0% opacity for /p/ (i.e., an unambiguous /t/ token) to 100% opacity for /p/ (i.e., an unambiguous /p/ token) by increasing the opacity for /p/ in increments of 5%.

c. Audiovisual exposure materials

Audiovisual exposure items consisted of eight natural target words ending in /p/ or /t/ and eight natural nonsense words ending in /f/ or /x/. In addition, eight ambiguous versions of these target words were created with auditorily and visually ambiguous final plosives. To create the visually ambiguous plosives, we selected two audiovisual tokens for each target word (i.e., the target word and the same word ending in the alternative plosive) on the basis of how well they could be merged

visually. For each of the eight target words a visual-only and auditory-only continuum was created using the same stimulus creation procedures detailed for the auditory and visual pretest materials. The most ambiguous visual step for each target word ($V_?$) was established on the basis of a pilot study and was the same across participants but different across target words. The video containing this step was combined with an audio track containing each participant's most ambiguous auditory step ($A_?$), as found in the auditory-only pretest for each participant. This created target words in which the critical sounds were ambiguous in both modalities ($A_?V_?$).

d. Visual-only pilot

A pilot study was conducted to test participants' sensitivity to the visual-only *soop-soot* continuum and to select the most ambiguous visual continuum step for each of the eight target words. Participants categorised 13 steps from the *soop-soot* continuum (steps 0, 15, 30, 35, 40, 45, 50, 55, 60, 65, 70, 85, 100). Participants also categorised 10 steps (steps 0, 15, 30, 35, 40, 45, 50, 55, 60, 65, 70, 85, 100) from four of the eight target word continua. The four target word continua always consisted of two /p/-final targets and two /t/-final targets, assigned randomly to each participant. The *soop-soot* continuum was always presented first. The presentation order of the following four target-word continua was rotated across lists. For every continuum, each step was repeated eight times in a newly randomised order within each repetition. The two response alternatives (i.e., /p/ or /t/) were displayed on a computer screen beneath the video of the speaker producing an utterance. Stimuli were presented 200 ms after trial onset. Participants were instructed to respond as accurately and as quickly as possible by pressing one of the two buttons on a button box that corresponded with the "p" and "t" labels shown on the computer screen. Each new trial started only after participants had given a response. No feedback was provided.

The results of the pilot study can be seen in Figure 1a and 1b. Figure 1a shows the results for the visual-only *soop-soot* continuum and Figure 1b the results for the visual-only target-word continua. The results indicate that participants were sensitive to the visual-only continua for both *soop-soot* and the target words and gave more [p] responses the more /p/-like the continuum step. The most ambiguous visual continuum step for each of the eight target words was selected on the basis of the 50% cut-off points, indicated by the vertical lines in Figure 1b. These steps were chosen as $V_?$ for the creation of the audiovisual exposure versions of these target words. Whenever the 50% point fell between two categorised steps, a new video was created with a step that was between the two steps adjacent to the 50% point. Four of the target stimuli contained such a newly created step (*kroop*, *zoop*, *goot*, and *schoot*). The selected steps for these target words were 52, 54, 51, and 43, respectively (cf. Figure 1b).

2.1.3. Design and procedure

Participants were randomly assigned to either the /p/-exposure group or the /t/-exposure group and tested individually in a sound-attenuated booth. The experimental session lasted 45 minutes. Participants started the experiment with an auditory-only pretest in which they categorised 15 steps from the auditory-only *soop-soot* continuum (steps 1, 4, 6-16, 18, 21). All continuum steps were presented eight times in a newly randomised order for every repetition. The audio was presented over Sennheiser HD280 headphones at a fixed level. Participants indicated whether the final sound they had heard was /p/ or /t/ by clicking with the computer mouse on labelled buttons on a computer screen. Each new trial started 500 ms after a response had been given. The results for the auditory-only pretest were used to select each participant's most ambiguous auditory token $A_?$ for use in the rest of the experiment. $A_?$ was always the step closest to participants' 50% cut-off point between [p] and [t].

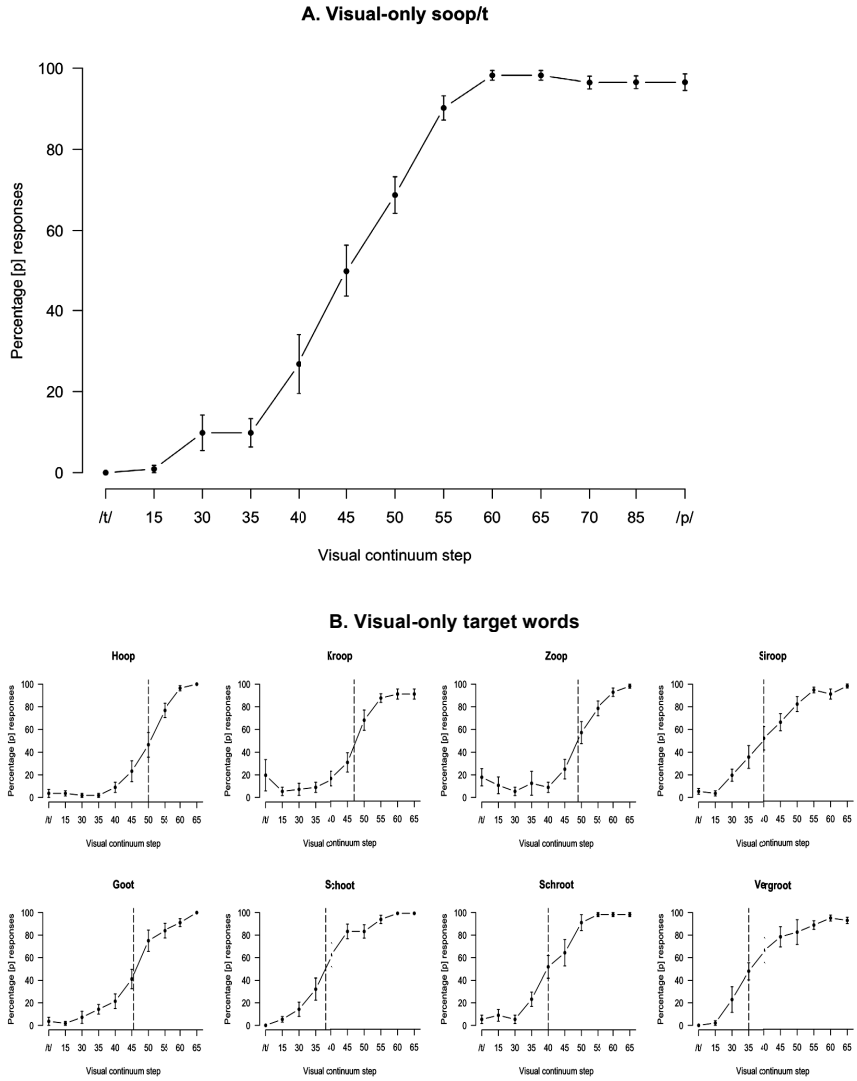

Figure 1: Mean percentages of [p] responses as a function of /so:p/-/so:t/ continuum steps (Panel A) and for the visual-only continua of all eight target words (Panel B) in the visual-only pilot study. Horizontal lines mark 50 percent [p] responses. Vertical lines mark the visual step used to create the audiovisual exposure materials. Error bars show the standard error of the mean.

After the auditory-only pretest, participants performed a visual-only pretest. Participants categorised seven steps from the visual-only *soop-soot* continuum (steps 0, 35, 40, 45, 50, 55, 100). Each step was presented three times with presentation

blocked by repetition. Participants indicated whether the final sound the talker had produced was a /p/ or a /t/ by pressing the button on a button box that corresponded to the respective labels shown on-screen. New trials started 800 ms after participants gave a response. This visual-only pretest provided a baseline to which the posttest results were compared.

The exposure phase consisted of an audiovisual lexical decision task. Each exposure block was immediately followed by another visual-only categorisation block (posttest) and participants completed a total of 10 repetitions of such exposure-posttest sequences. Participants received four /t/-final and four /p/-final target words, intermixed with four /f/-final and four /x/-final nonsense words in each exposure block. Participants assigned to the /p/-exposure group received /p/-final target words where the final plosive was both visually and auditorily ambiguous ($A_?V_?$) along with natural /t/-final target words (A_tV_t). Participants in the /t/-exposure group received auditorily and visually ambiguous /t/-final words ($A_?V_?$) along with natural /p/-final words (A_pV_p). The exposure condition was the same for a participant across all repetitions of the exposure and posttest phases. $A_?$ in the audiovisual exposure materials was selected on the basis of each participant's pretest results and the same in all words. $V_?$ in the materials was selected based on the pilot study data and the same for all participants in a given word, but different across words. Participants watched and heard the speaker produce each item and indicated as quickly and as accurately as possible whether or not what the talker had said was an existing Dutch word. Answers were provided by pressing the button on a button box that corresponded with the respective label shown on the computer screen ("w" for "wel"/"yes"; "n" for "niet"/"no"). All 16 items were presented twice in random order blocked by repetition. New trials started 800 ms after the participant gave a response.

2.2. Results and discussion

Results were analysed using linear mixed-effect models in the R statistical program (Version 2.11.0; R Development Core Team, 2007) by using the lmer function of the lme4 library (Bates & Sarkar, 2007). The dependent variable for the exposure phase was the binomial word judgment (correct or incorrect). The dependent variables for the pretest and posttests were the binomial response to the continuum steps (0 = /t/; 1 = /p/). A logistic linking function was used for these categorical dependent variables. The best-fitting model for each data set was established through systematic model comparison using likelihood-ratio tests. We always started with the full model, gradually removing factors that did not contribute to a better model fit, starting with the factors with the largest *p* values. Main effects were only removed if their factors did not contribute to an interaction. All best-fitting models included participants as a random factor. Group (/p/-exposure group vs. /t/-exposure group) was evaluated as a contrast-coded fixed factor in all analyses. Ambiguity (natural target words vs. ambiguous target words) was evaluated as a contrast-coded fixed factor in the analysis of the exposure data. Visual continuum step was evaluated as a numerical factor cantered on the middle step in the pretest and the posttest analyses. Test (pretest vs. posttest) was evaluated as a contrast-coded fixed factor in the comparison of the visual-only pretest and posttest data.

Table 1. *Mean Percentage Correct Responses to Natural and Ambiguous /p/-final and /t/-final Words in Experiment 1 and 2.*

	Natural		Ambiguous	
	/p/ words	/t/ words	/p/ words	/t/ words
Experiment 1	95.44	94.44	87.50	93.06
Experiment 2	92.45	96.30	81.76	95.31

2.2.1. Visual-only pretest

There was no difference in the number of [p] responses given by the two groups at pretest (not a predictor, β = -0.31, standard error (*SE)* = 0.48, p = .52). Both groups gave more [p] responses to the more /p/-like visual tokens (β = 0.20, *SE* = 0.01, p < .001; see Figure 2). This indicates that the two groups were sensitive to the visual-only continuum and did not differ prior to testing in their visual categories.

2.2.2. Audiovisual exposure

Table 1 (upper row) gives the mean percentages of correct 'word' responses to ambiguous and nonambiguous versions of the target words. Participants gave more correct responses to the natural target words than to the target words containing an ambiguous plosive (β = 0.77, *SE* = 0.13, p < .001). This difference between natural and ambiguous target words was numerically larger in the /p/-exposure group (natural: 94%, ambiguous: 88%) than in the /t/-exposure group (natural: 95%, ambiguous: 93%), but the interaction was only marginally significant ($\chi^2(1) = 3.58$, p = .06).

2.2.3. Visual-only posttests

The data from all visual-only posttest blocks were pooled together since there was no effect of block (β = -0.00, *SE* = 0.01, p = .96). Participants gave more [p] responses to the more /p/-like visual continuum steps in the posttest, again indicating sensitivity to the visual-only continuum (β = 0.19, *SE* = 0.01, p < .001). Participants in the /p/-exposure group gave more [p] responses than participants in the /t/-exposure group (β = -1.21, *SE* = 0.50, p < .05). This result indicates an effect of learning in line with exposure. Lexical knowledge can thus be used to retune visual phonetic categories. Participants in the /p/-exposure group gave more [p] responses in the posttests than in the pretest (β = 0.92, *SE* = 0.19, p < .001). The responses from participants in the /t/-exposure group in the posttests did not differ from the pretest

($\chi^2(1) = 1.49$, $p = .22$). This indicates that, while there is a difference between the two groups in line with their exposure, this difference between the groups is mainly due to learning in the /p/-exposure group.

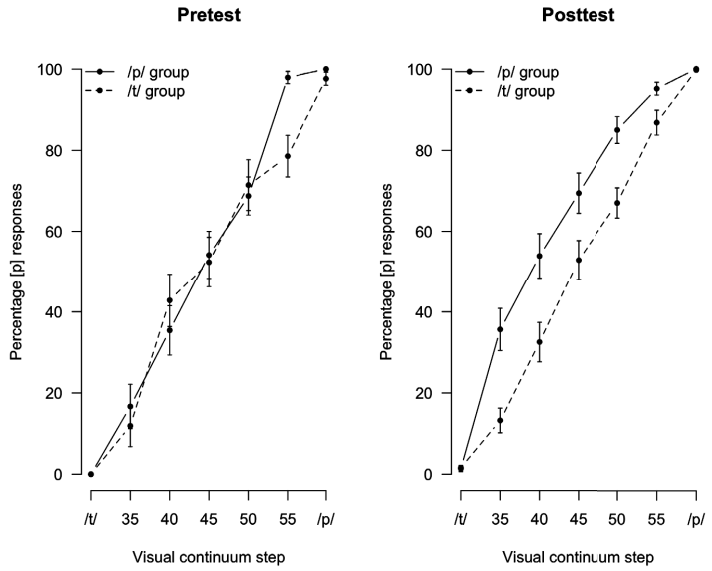

Figure 2: Mean percentages of [p] responses across pretest and posttests as a function of visual continuum step in Experiment 1. Solid lines show the results for the /p/-exposure group and dashed lines the results for the /t/-exposure group. Error bars show the standard error of the mean.

3. Experiment 2

In Experiment 1, we showed that lexical knowledge could be used to shift the boundaries of visual phonetic categories. Exposure to an audiovisually ambiguous sound within a biasing lexical context resulted in a shift of the visual category boundary. This shift was only observed for the /p/-exposure group, but not for the /t/-exposure group. Listeners in Experiment 1 could either have used lexical knowledge to retune visual phonetic categories directly, or used lexical information to retune auditory category boundaries, which in turn influenced visual category

boundaries. The observed shift for the visual category boundaries could in the latter case reveal generalisation across modalities. In Experiment 2, we directly tested whether retuning of the visual phonetic categories can occur through generalisation of speaker knowledge across modalities. In Experiment 2, participants were exposed to auditory-only versions of the audiovisual stimuli of Experiment 1 and were subsequently tested on the visual-only continuum and on an auditory-only version of that continuum. This way, we investigated whether retuning of visual categories can still occur even when visual speech was not presented with the lexically disambiguating context.

3.1. Methods

3.1.1. Participants

Forty-four new participants (average age 20.8 years; 12 males) from the same population as for Experiment 1 were tested. Five participants were excluded due to insensitivity to the auditory continuum during the pretest. An additional eight participants were excluded for failing to exceed a threshold of 50 percent correct 'word' responses to the ambiguous target words on the lexical decision task. All of these excluded participants had been assigned to the /p/-exposure group. The final data set consisted of data from 31 participants, from 15 in the /p/-exposure group and from 16 in the /t/-exposure group.

3.1.2. Materials

Materials for Experiment 2 were the same as those used in Experiment 1. However, rather than audiovisual stimuli, participants received auditory-only versions of the stimuli during the exposure phase. The auditory-only stimuli were created by blacking out the video of the audiovisual stimuli used during exposure in Experiment 1. Stimuli were otherwise identical. The auditory-only posttest stimuli

were a subset of the steps of the auditory-only /so:p/-/so:t/ continuum used in the pretest.

3.1.3. Design and procedure

There were two differences between the procedure of Experiment 1 and Experiment 2. In Experiment 2, the exposure materials were auditory-only rather than audiovisual, and participants performed an additional auditory-only posttest at the end of the experiment. Otherwise the procedure of Experiment 2 was the same as in Experiment 1. First an auditory-only pretest established each participant's most ambiguous auditory step ($A_?$) for exposure. Participants then completed 10 exposure-posttest repetitions where they first performed an auditory-only lexical decision task (exposure) and then a visual-only categorisation task (posttest). After these exposure-posttest repetitions, participants completed an additional auditory-only categorisation task. This auditory test was added as a control to test whether the exposure materials would lead to retuning of auditory phonetic categories. It was conducted at the end of testing to ensure comparability between the visual-only posttest results for Experiments 1 and 2.

The auditory-only posttest consisted of three steps from the auditory-only *soop-soot* continuum, namely the participant's most ambiguous step $A_?$ and a more /p/-like step $A_{?-1}$ and a more /t/-like step $A_{?+1}$. All three steps were presented eight times in a newly randomised order for each repetition. Participants responded by pressing one of the buttons on a button box that corresponded to the labels shown on the computer screen.

3.2. Results and discussion

Results were analysed as for Experiment 1. Group (/p/-exposure group vs. /t/-exposure group) was evaluated as a contrast-coded fixed factor and auditory continuum step as a fixed factor cantered on the middle step in the analysis of the

auditory-only posttest data. Participants were included as a random factor in the best-fitting model for the auditory-only posttest.

3.2.1. Visual-only pretest

The two groups did not differ in the number of [p] responses given in the visual-only continuum steps at pretest (not a predictor, $\chi^2(1) = 0.10$, $p = .75$). Both groups were sensitive to the visual-only continuum and gave more [p] responses the more /p/-like the visual continuum steps were ($\beta = 0.18$, $SE = 0.02$, $p < .001$). This indicates that the two groups were sensitive to the visual-only continuum but their visual categories did not differ prior to exposure.

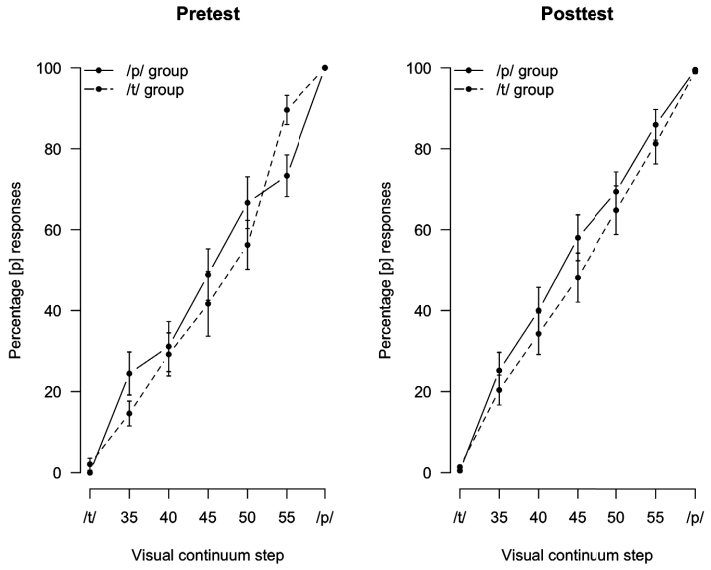

Figure 3: Mean percentages of [p] responses across pretest and posttests as a function of visual continuum step in Experiment 2. Solid lines show the results for the /p/-exposure group and dashed lines the results for the /t/-exposure group. Error bars show standard error of the mean.

3.2.2. Auditory-only exposure

There was no difference between the responses of the /p/-exposure group and the /t/-exposure group in the exposure phase (not a predictor, β = 0.44, *SE* = 0.52, p = .39; see Table 1, lower row). Overall, participants gave more correct responses to the natural target words than to the ambiguous target words (β = 0.77, *SE* = 0.13, p < .001). The difference between responses to the natural and ambiguous target words for the /p/-exposure group (natural: 96%; ambiguous 82%) was opposite to that observed for the /t/-exposure group (natural 92%: ambiguous 95%; β = -2.55, *SE* = 0.27, p < .001). The /p/-exposure group gave more correct responses to the natural target words than to the ambiguous target words (β = 1.98, *SE* = 0.19, p < .001), while the /t/-exposure group gave fewer correct responses to the natural target words than to the ambiguous target words (β = -0.57, *SE* = 0.19, p < .01). The unexpected pattern for the /t/-exposure group may have been due to the unambiguous item *zoop*, which had been rejected as a word in 42% of all presentations. This item may have been categorised as a nonword, since participants may have thought of it as being too colloquial or dialectal to be a real Dutch word.

*3.3.3. Visual-only posttest*s

The results from the visual-only posttest revealed no differences between the number of [p] responses given by the two groups (not a predictor, $\chi^2(1) = 0.51$, p = .47), indicating that auditory-only exposure did not affect the subsequent categorisation of the visual-only continuum (see Figure 3). Participants did thus not retune their visual category boundaries after auditory-only exposure. Participants in both groups were sensitive to the visual-only continuum and gave more [p] responses the more /p/-like the continuum step was (β = 0.18, *SE* = 0.01, p < .001).

3.3.4. Auditory-only posttest

Overall, participants were sensitive to the auditory-only continuum and gave more [p] responses to the more /p/-like steps (β = 0.50, SE = 0.11, $p < .001$; see Figure 4). Participants in the /p/-exposure group gave more [p] responses than those in the /t/-exposure group (β = -1.38, SE = 0.56, $p < .05$), indicating that the categorisation of the auditory-only posttest was influenced by exposure. This finding replicates results reported by earlier studies by showing that lexical information can guide retuning of auditory phonetic categories (McQueen, Cutler et al., 2006; McQueen, Norris, & Cutler, 2006; Norris et al., 2003). Taken together, the results of the auditory-only posttest and the visual-only posttests show that while listeners used lexical information here to retune their auditory phonetic categories based on the auditory-only exposure, this retuning did not affect visual phonetic categories.

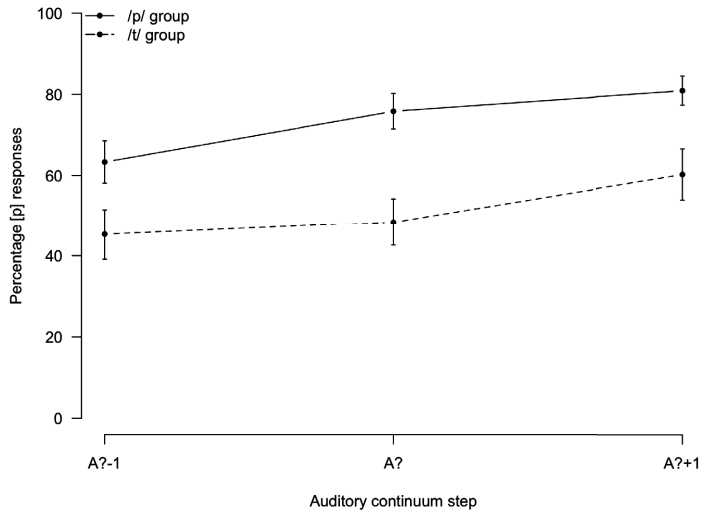

Figure 4: Mean percentages of [p] responses for the auditory-only posttest in Experiment 2 as a function of auditory continuum step. Solid lines show the results for the /p/-exposure group and dashed lines the results for the /t/-exposure group. Error bars show the standard error of the mean.

4. General discussion

Listeners perceive speech bimodally when they hear and see someone talk. Idiosyncrasies of a speaker expressed in one modality can be disambiguated by information in the simultaneously presented speech in the other modality (Baart & Vroomen, 2010; Bertelson et al., 2003). This disambiguation leads to the retuning of category boundaries in line with the disambiguating context. Listeners also use their lexical knowledge to retune auditory phonetic categories to talker idiosyncrasies contained in auditory speech (Norris et al., 2003). The results of the present study show that lexical knowledge can also retune visual phonetic categories. Exposure to audiovisually ambiguous sounds that were disambiguated by lexical information resulted in shifts of listeners' visual category boundaries. Furthermore, the current results also indicate that visual phonetic categories are only influenced by lexical knowledge, when visual information about the idiosyncrasy was available to the listener. Auditory-only exposure to an idiosyncratic sound resulted in retuning of auditory phonetic categories but did not affect visual phonetic categories. Phonetic retuning in one modality does not generalise to the categories in another modality.

Listeners use their lexical knowledge to adjust visual category boundaries to optimise speech recognition. Retuning the visual phonetic categories in this way is particularly beneficial in situations where the same idiosyncrasy is observed in both the auditory and the visual modality. In such cases, information from neither modality can be used to guide the perceptual learning. Listeners are then dependent on other sources, such as their linguistic knowledge, for the resolution of the ambiguity in the audiovisual speech input. Listeners use lexical knowledge to directly retune their visual phonetic categories, or do so indirectly via the retuning of auditory categories. Our results show, however, that listeners were only able to adjust their visual category boundaries if the lexicon disambiguated the visual idiosyncrasy as /p/. This could indicate that retuning of the category boundaries only occurs for those phonemes that are strongly defined visually (here, the bilabial plosives), but not for those phonemes that are difficult to identify visually (here, the

alveolars, for which the defining place of articulation is inside the oral cavity). That is, a departure from typicality that is not readily noticeable to the eye will not prompt category retuning. Although only further research will conclusively decide the issue, listeners may only be sensitive to speaker idiosyncrasies in phonemes that are visually distinct, and in consequence it may be only such phoneme categories that are retuned.

It should be noted that the ability to resolve visual ambiguity by reference to existing knowledge, and to apply learning from such ambiguity resolution to future visual perceptual processing, is by no means confined to speech recognition. The interpretation of colour in visual processing involves similar perceptual learning operations, as a colour-perception analogue of the Norris et al. (2003) experiment showed. Mitterer and De Ruiter (2008) presented viewers with pictures of fruit, typically encountered either in yellow or orange, in an ambiguous colour between yellow and orange, and then collected categorisation judgments on a yellow-orange continuum of coloured socks. Viewers who had seen the ambiguous colour on bananas judged more socks along the continuum as yellow, whereas viewers who had seen the ambiguous colour on oranges categorised more socks as orange. The same kind of visual category shift was also observed with ambiguous letters between H and N presented word-finally in sequences such as WEIG- versus REIG- (Norris, Butterfield, McQueen, & Cutler, 2006). In our complex world, sensory processing in any modality is liable to deliver ambiguous input, but our cognitive processing is able to resolve the ambiguity by referring to knowledge of many sorts, and can learn from this to improve future processing.

The results of Experiment 2 provide evidence that the visual phonetic categories were only influenced by listeners' lexical knowledge if visual information about the speaker's idiosyncrasy was available to the listener. Phonetic retuning occurred for listeners' auditory phonetic categories after exposure to auditory-only idiosyncratic speech, but no such retuning was observed for the visual phonetic categories. Lexically guided retuning in one modality thus did not generalise to

another modality and the boundary shifts for the visual phonetic categories in Experiment 1 must have occurred, because listeners obtained information about how to retune their visual categories directly from seeing the speaker talk. For retuning to occur, information about the idiosyncrasy needs to be available to the listener from the modality for which the phonetic categories are retuned.

Transfer for speaker information across modalities has been observed in a previous study, however (Rosenblum et al., 2007). Rosenblum and colleagues found transfer of knowledge about a speaker's visual speech to their auditory speech. A variety of methodological differences between the Rosenblum study and the current study could provide an explanation for the discrepancy in the findings. Most notably, participants in the Rosenblum study received the critical words in sentences during exposure and test. In our study, participants were presented with isolated words during exposure and nonsense syllables during test. Words are generally more easily identified when presented in a meaningful sentence context than when they are presented in isolation (Boothroyd & Nittrouer, 1988; Grant & Seitz, 2000; Miller, Heise, & Lichten, 1951). This, in addition to the increased amount of exposure in the Rosenblum study compared to our study, could have lead to better learning and therefore cross-modal transfer of speaker information. But because words were presented in sentences, listeners in the Rosenblum study could also arguably have been familiarised with, and subsequently have generalised, different properties of the speaker than listeners in our study. Speaker familiarity established on the basis of sentences does not significantly improve subsequent recognition of novel words in isolation (Nygaard & Pisoni, 1998), indicating that listeners may tune in to a different set of speaker-specific properties depending on the exposure materials. Sentences provide information about speaker-specific properties such as prosody, duration and speaking rate (Adank & Janse, 2009; Grant et al., 1998; Nygaard & Pisoni, 1998), to which listeners can attune, but which are not available from isolated words. Learning of these speaker characteristics could possibly transfer across modalities (see, for

instance, Cvejic, Kim, & Davis, 2012), while learning of phonetic idiosyncrasies, as tested in our study, may not.

Retuning for auditory and visual phonetic categories thus appears to reflect two distinct processes that do not necessarily affect one another. Listeners retune their boundaries for whichever category is problematic during exposure to a speaker, considering all available information. If speech from only one modality is provided, then only the boundaries of categories for that modality are changed and this shift does not affect the category in the other modality. Retuning for the visual category failed in Experiment 2, because the ambiguity was only presented in the auditory modality and so listeners were not aware of how to retune their visual category boundary. This finding indicates that auditory and visual categories are not inextricably linked and that changes for the categories in one modality do not necessarily result in changes for the categories in the other modality.

The results of Experiment 2 pose a potential problem for theories that posit that listeners use information about the speaker's intended vocal tract gestures for speech perception, i.e., motor theory and direct realist theory (Fowler, 1986, 1991; Fowler, Brown, Sabadini, & Welhing, 2003; Galantucci, Fowler, & Turvey, 2006; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985). In these theories, it is postulated that listeners are able to obtain information about the underlying gestures from auditory speech input. If this were the case, then listeners should be able to retune their visual phonetic categories based on auditory speech alone. That is, if lexical knowledge disambiguates an auditory speaker idiosyncrasy, then the auditory speech signal alone should contain all the information necessary to retune the characteristic articulatory features that encompass the corresponding visual phonetic category. The finding that lexically guided retuning of auditory categories does not transfer to visual categories in Experiment 2 suggests, however, that such information about the articulatory movements is not extracted (or directly perceived) from the auditory speech input.

Instead, auditory-only presentation results in boundary shifts only for auditory phonetic categories.

In the present experiments, we have shown that reference to information outside the speech signal itself is deployed for visual as for auditory ambiguity resolution. Such information can be lexical, as in the present experiments and in many others, but it need not be; for instance, phonotactic constraints realised in nonword sequences also lead to similar learning (Cutler, McQueen, Butterfield, & Norris, 2008). Our study indicates that while there is a tight link between auditory and visual speech, the respective categories are separate and retuning of each is a separate process.

5. Conclusions

The present study extends our knowledge about lexically guided retuning of phonetic categories. First, we have demonstrated that lexical information can guide retuning of visual phonetic categories. Second, lexical information does not retune visual categories through generalisation across modalities. Despite the inherent link between auditory and visual speech, listeners do not adjust their visual category boundaries on the basis of lexically retuned auditory category boundaries. Retuning based on lexical information helps learning about the idiosyncrasies in the modality they occur in but does not generalise across modalities.

Chapter 3:

Cross-speaker generalisation in two phoneme-level perceptual adaptation processes

Van der Zande, P., Jesse, A., & Cutler, A. (Under revision). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics.*

Abstract

Speech perception is shaped by listeners' prior experience with speakers. Listeners retune their phonetic category boundaries after encountering ambiguous sounds in order to deal with variations between speakers. Repeated exposure to an unambiguous sound, on the other hand, leads to a decrease in sensitivity to the features of that particular sound. This study investigated whether these changes in the listeners' perceptual systems can generalise to the perception of speech from a novel speaker. Specifically, the experiments looked at whether visual information about the identity of the speaker could prevent generalisation from occurring. In Experiment 1, listeners retuned auditory category boundaries using audiovisual speech input. This shift in the category boundaries affected perception of speech from both the exposure speaker and a novel speaker. In Experiment 2, listeners were repeatedly exposed to unambiguous speech either auditorily or audiovisually, leading to a decrease in sensitivity to the features of the exposure sound. Here, too, the changes affected the perception of both the exposure speaker and the novel speaker. Together, these results indicate that changes in the perceptual system can affect the perception of speech from a novel speaker and that visual speaker identity information did not prevent this generalisation.

1. Introduction

The speech we encounter in our everyday communication is highly variable. The speech perception system of the listener is flexible, however, and capable of dealing with this variation. In fact, the perceptual system is continually adjusted following the input with which it is provided. Phonetic retuning and selective adaptation are two distinct adaptation processes that show how effectively the perceptual system can be adjusted on the basis of speech input (Bertelson, Vroomen, & De Gelder, 2003; Diehl, 1975; Eimas & Corbit, 1973; Norris, McQueen, & Cutler, 2003). In the current experiment, we investigated the generality of these two adaptation processes. We specifically looked at whether the changes that occur within the perceptual system affect the subsequent perception of only the speech produced by the speaker to whom the system adjusted or whether speech perception for different speakers is also affected.

Since the perceptual system is flexible, it can adapt to many different subtle features of the speech input. Adjustments within the perceptual system occur for speech that is unambiguous and clearly intelligible and for speech that is somehow problematic. Adaptation caused by unambiguous speech, as seen with selective adaptation, may reflect overexposure to a particular sound, while adaptation to ambiguous sounds shows how the perceptual system deals with problematic input. Listeners are able to adjust to variability in the input, for instance, when encountering speech produced by non-native speakers of the language (Bradlow & Bent, 2008; Clarke & Garrett, 2004) or by speakers with a distinct accent (Maye, Aslin, & Tanenhaus, 2008). Adaptation also occurs for native speech input since the realisation of sounds is idiosyncratic to each individual speaker. Speech is generally more accurately identified when produced by a familiar speaker than when it is produced by a speaker that we have not previously encountered (Bradlow, Nygaard, & Pisoni, 1999; Craik & Kirsner, 1974; Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). This effect of speaker familiarity is due to the fact that speech perception is facilitated when listeners have become attuned to

speaker-specific idiosyncrasies, for instance through the process of phonetic retuning (Baart & Vroomen, 2010; Bertelson et al., 2003; Norris et al., 2003).

Speaker-specific idiosyncrasies can render sounds ambiguous but listeners are able to disambiguate speech by referring to their stored knowledge about the language (Norris et al., 2003). When listeners hear the word *platypu*[?] (where [?] symbolises an ambiguous sound between /f/ and /s/) they are still able to correctly identify the word despite the ambiguity in the auditory input. Listeners can use lexicl information to help disambiguate the sound because their knowledge of English tells them that *platypus* is a word but *platypuf* is not. In situations where lexical information cannot help to disambiguate the sound, listeners can also use visual speech input because this provides information that is redundant and complementary to the available auditory information (Grant, Walden, & Seitz, 1998; Jesse & Massaro, 2010; Sumby & Pollack, 1954; Summerfield, 1987; Walden, Prosek, & Worthington, 1974). Exposure to ambiguous speech causes shifts in listeners' phonetic category boundaries and these shifts occur in order to assign the ambiguous input to the appropriate category as dictated by the disambiguating source of information.

Changes in the category boundaries affect how listeners subsequently judge the ambiguous sounds. Hearing [?] in the context of *platypus* makes listeners give more [s] responses to steps of an /f/-/s/ continuum, while hearing the same sound in the context of *giraffe* has the opposite effect (Norris et al., 2003). Disambiguation by the visual speech signal can also cause shifts in the auditory phonetic categories (Bertelson et al., 2003). Phonetic retuning thus facilitates the subsequent recognition of sounds and does so across the full extent of the lexicon, even when the word context or the word-internal position is changed (Jesse & McQueen, 2011; McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011; Sjerps & McQueen, 2010).

Adaptation does not occur only with difficult-to-process input, however; changes in the perceptual system are also made after exposure to unambiguous speech input. Repeated exposure to an unambiguous sound leads to a decrease in

sensitivity to the features of that particular sound (Diehl, 1975; Eimas & Corbit, 1973; Samuel, 1986; Sawusch, 1977; Sawusch & Pisoni, 1976). This reduced sensitivity to specific phonetic features is thought to be due to fatigue within the perceptual system (Samuel, 1986). Like phonetic retuning, selective adaptation affects listeners' perception of sounds but does so in the opposite direction (Eimas & Corbit, 1973). Listeners give fewer [da] responses to the steps of a /ba/-/da/ continuum after multiple repetitions of an unambiguous /da/ utterance than after multiple repetitions of /ba/. Phonetic retuning and selective adaptation thus clearly reflect distinct processes within the perceptual system.

Selective adaptation to auditory features has been shown to be based on acoustic information and not on the perceived identity of the speech input (Blumstein, Stevens, & Nigro, 1977; Sawusch & Pisoni, 1976). Whereas phonetic retuning can be guided by visual speech input, selective adaptation is not modulated by visual speech information. To show this, one can take advantage of the McGurk effect (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). Listeners presented with an auditory /ba/ accompanied by seeing a speaker produce /ga/ perceive this audiovisual stimulus as /da/, showing the influence of the visual speech information on the perception of the auditory input. After repeated exposure to a similar incongruent but perceptually unambiguous stimulus, listeners show adaptation to the sound they were presented with auditorily (/ba/) rather than to what they had perceived, namely /va/ (Saldaña & Rosenblum, 1994). The fact that selective adaptation is in line with the acoustic signal even when this differs from the perceived identity of the audiovisual utterance suggests that selective adaptation is modality-specific and takes place before the auditory and visual speech signals are integrated (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994).

The integration of information from both speech modalities appears not to be necessary for selective adaptation but it is for phonetic retuning, as shown in a study using sine-wave speech (Vroomen & Baart, 2009a). Sine-wave signals are stripped of much of the acoustic detail of speech but retain overall amplitude and frequency

cues. Listeners generally do not perceive sine-wave speech as containing speech information until they are explicitly informed. Listeners in Vroomen and Baart's study were exposed to sine-wave speech combined with simultaneously presented visual speech. Integration of the two speech signals only happened when listeners were aware of the speech origins of the sine-wave input. Effects of selective adaptation were observed regardless of whether listeners were informed about the sine-wave speech signal and thus regardless of whether intersensory integration had taken place. Phonetic retuning, on the other hand, was only observed for informed listeners and thus appears to be dependent on the integration of the auditory and the visual speech input.

A final piece of evidence for the dissociation of the two effects is provided by the difference in the rates at which they build up and dissipate (Vroomen, Van Linden, De Gelder, & Bertelson, 2007; Vroomen, Van Linden, Keetels, De Gelder, & Bertelson, 2004). Selective adaptation builds up slowly and remains for up to 60 consecutive categorisation trials without renewed exposure (Vroomen et al., 2004). The slow build-up suggests that it takes some time for fatigue within the perceptual system to set in. Phonetic retuning, on the other hand, is established rapidly with only a small number of exposure trials (Vroomen et al., 2007), indicating that learning occurs nearly instantly after perceiving problematic speech input. The rate of dissipation for phonetic retuning varies depending on the source of the disambiguating information during exposure. Visually guided retuning dissipates quickly and the effect is no longer observed after six categorisation trials, unless there is additional exposure (Vroomen & Baart, 2009b; Vroomen et al., 2004). The effects of lexically guided retuning are still observed after a 25-minute or even a 12-hour intervening period between exposure and test (Eisner & McQueen, 2006; Kraljic & Samuel, 2005), although the studies investigating visually guided and lexically guided retuning varied on more points than just the source of the disambiguating information.

Given the fact that phonetic retuning and selective adaptation are different processes of adaptation, they may also differ in the extent to which their influence affects speech from a novel speaker. Adjustments made for one speaker could potentially be applied to the perception of speech from another speaker. On the other hand, adjustments in the perceptual system could also be speaker specific. Selective adaptation has been shown to generalise across phonemes (Eimas & Corbit, 1973) but not across syllable positions (Ades, 1974). Generalisation of selective adaptation across speakers has been found, however, with static visual representations of speech sounds (Jones, Feinberg, Bestelmeyer, DeBruine, & Little, 2010). Exposure to still images of a speaker producing a sustained /m/ sound resulted in fewer [m] responses than exposure to an image of the speaker producing a sustained /u/ sound when images showing mouth shapes ambiguous between /m/ and /u/ were subsequently categorised. The same effect of selective adaptation to the mouth shapes was observed for the exposure speaker and for a novel speaker. It remains unclear, however, whether selective adaptation to auditory speech also generalises across speakers following exposure to natural auditory and audiovisual speech materials.

Unlike selective adaptation, phonetic retuning has already been shown to generalise across speakers. More specifically, lexically guided retuning generalises across speakers when the critical phonemes are plosives, but not when they are fricatives (Eisner & McQueen, 2005; Kraljic & Samuel, 2006). Exposure to an ambiguous sound between /d/ and /t/ resulted in effects of phonetic retuning after exposure regardless of whether the subsequently categorised speech was produced by the exposure speaker or by a novel speaker (Kraljic & Samuel, 2006). Generalisation across speakers was not found after exposure to an ambiguous fricative between /f/ and /s/, however (Eisner & McQueen, 2005).

The discrepancy between these findings has been attributed to differences in the phoneme contrasts that were used (Kraljic & Samuel, 2006). The voicing distinction for the plosive sounds depends, among others, on the duration of the

silence before the release and the duration of vibration after the release (in both cases longer durations favour /t/). These durational cues occur on a single dimension, so while speakers may vary in their durations (Allen, Miller, & DeSteno, 2003), the nature and the direction of the effect is constant, making learning for one speaker applicable to the recognition of speech from other speakers (Kraljic & Samuel, 2005). The place distinction for fricatives is based on spectral cues, which depend on the shape of the speaker's vocal tract and vary more substantially across speakers. This variability makes learning for fricatives specific to individual speakers and it does not generalise (Eisner & McQueen, 2005; Kraljic & Samuel, 2005).

Generalisation across speakers may thus be driven by the acoustic similarity for the target phonemes across speakers. An alternative explanation might be that generalisation is influenced by the availability of speaker identity information in the input. The results of the previous studies cannot speak to which of these two factors matters, since the degree of acoustic similarity across speakers and the degree to which the speech sounds contained speaker identity information were confounded. In the current study, we investigated this problem directly by teasing apart the acoustic similarity and the availability of speaker identity information. To do so, we used audiovisual speech materials in combination with a plosive contrast. We used two plosive sounds (/b/ and /d/) in order to provide a favourable auditory context for generalisation to occur. Place of articulation was used rather than a voicing contrast, because the former but not the latter can be distinguished by listeners on the basis of visual speech (Bernstein, Demorest, & Tucker, 2000; Van Son, Huiskamp, Bosman, & Smoorenburg, 1994).

For phonetic retuning, the focus of the current study was to determine whether generalisation takes place after exposure to audiovisual speech. The auditory speech input should allow generalisation while the visual speech input contains information about the identity of the speaker, which may inhibit generalisation. In Experiment 1, participants were presented during exposure with audiovisual speech tokens containing an auditory ambiguity that was resolved by

the visual speech signal. Exposure to such audiovisual materials should induce phonetic retuning. A subsequent auditory-only test phase had participants categorise continua steps produced by either the exposure speaker or by a novel speaker. If acoustic similarity drives phonetic retuning, we should see an effect of retuning for both speakers at test. If generalisation is affected by speaker identity information, however, only a diminished effect or no effect of generalisation is expected. In Experiment 2, the possibility of generalisation across speakers for selective adaptation was investigated in both an auditory-only and an audiovisual condition. Participants received unambiguous auditory and audiovisual speech materials during exposure, sufficient to induce selective adaptation. Since selective adaptation has been shown to be unaffected by visual speech input, generalisation across speakers is expected for both presentation conditions. If, on the other hand, information about the identity of the speaker in the visual speech input does affect generalisation, generalisation should only be observed in the auditory-only condition.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty-eight native speakers of Dutch (mean age = 20; 8 males) were paid for their participation in Experiment 1. All participants reported having normal hearing and normal or corrected-to-normal vision. Three participants were excluded due to equipment failure. One further participant was excluded due to insensitivity to the auditory continuum in the calibration phase. The final data set used for analysis consisted of the data from 24 participants. Seven additional participants from the same population took part in an auditory-only pilot experiment.

2.1.2. Materials

Two male native speakers of Dutch were video recorded with a Sony DCR-HC1000E camera. Audio was recorded simultaneously with two stand-alone Sennheiser microphones. Videos showed the head and shoulders of a speaker. The recordings of the two talkers formed the basis for all materials used in Experiment 1 and in Experiment 2. Talkers produced multiple tokens of the nonsense vowel-consonant-vowel (VCV) utterances /aːba/, /aːda/, and /aːxa/. These utterances were produced in pairs, avoiding list intonation. All possible combinations of the CVC tokens were recorded. Videos were digitised as uncompressed avi files (720 × 576 pixels in PAL format). Audio sampling rate was 44.1 kHz.

a. Auditory-only test materials

For both speakers an individual auditory-only /aːba/-/aːda/ continuum was created using Praat (Boersma & Weenink, 2006). Auditory /aːba/, /aːda/, and /aːxa/ tokens were selected; to avoid mismatched timing of features when combined with the visual speech input the selected tokens had feature durations as close as possible to the average durations across all recorded tokens of the same type. To create the continua, initial /aː/ sounds were first taken from the selected /aːxa/ tokens to ensure that the vowel transitions of the word-initial vowels did not contain any cues for either a following /b/ or a /d/. Parts of the steady-state portion of the initial /aː/s were removed so that the resulting sounds corresponded in duration to the average duration of /aː/ in this position across all tokens for the same speaker (approximately 265 ms and 375 ms for Speaker 1 and 2, respectively). Second, /ba/ and /da/ from the /aːba/ and /aːda/ tokens were edited to have equal durations and pitch contours before being mixed into a 21-step continua changing from /ba/ to /da/ in equal steps. These 21 steps were then concatenated with the edited initial

/aː/ token taken from /aːxa/ of the same speaker to create the final /aːba/-/aːda/ continuum.

A pilot study with seven participants was conducted in order to test participants' sensitivity to the two resulting continua. Participants categorised 13 continuum steps from both speakers' continua (steps 0, 3, 5, 6, 7-12, 13, 15, 17, 20). Each step was presented eight times in a newly randomised order within each repetition. The order of presentation of the two speakers was counterbalanced across participants. Continuum steps were presented over headphones at a fixed level. The response alternatives "b" and "d" were displayed on a computer screen and participants categorised the sounds by clicking on one of the two labels. Participants were instructed to respond as quickly and as accurately as possible. Each new trial started only after participants had given a response.

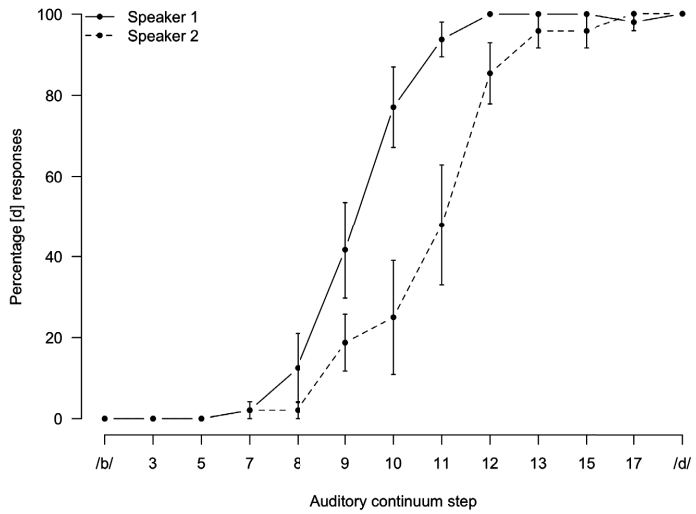

Figure 1: Mean percentages of [d] responses as a function of /aːba/-/aːda/ continuum steps in the auditory-only pilot. Solid lines show the results for Speaker 1 and dashed lines the results for Speaker 2.

Figure 1 shows the results of the pilot study for both speakers' auditory continua. The results indicate that the percentage of [d] responses increased the more /d/-like the auditory continuum step was and that participants were thus sensitive to the continua. These pilot results were used to select an ambiguous range of steps to be used in the main experiment. This range was between step 7 and step 13 of the continuum and was the same for both speakers.

b. Audiovisual exposure materials

The six steps that made up the ambiguous range for both speakers were combined with the natural visual speech tokens of /aːba/ and /aːda/ in order to create the audiovisual tokens $A_?V_b$ and $A_?V_d$. The visual-only speech tokens came from the same audiovisual tokens that provided the auditory speech input used for the auditory-only continua. Each audiovisual token started and ended with 15 frames showing the face of the speaker in a neutral position and with the lips parted slightly.

2.1.3. Design and procedure

Participants were tested individually in a sound-attenuated booth. The experiment consisted of three separate phases, similar to the design used by Bertelson and colleagues (2003). Participants first performed an auditory-only calibration phase before completing a number of exposure-test sequences where each of the 32 audiovisual exposure phases was directly followed by an auditory-only test phase.

The auditory-only calibration phase consisted of a phonetic categorisation task. The results of this categorisation task were used to select each participant's most ambiguous auditory continuum step ($A_?$). The selected step was then used in the audiovisual exposure materials, combined with unambiguous visual speech input, and in the auditory-only test materials. Participants categorised 13 steps from the auditory continua of both speakers (steps 0, 3, 5, 7-13, 15, 17, 20). Each step was

shown eight times in a newly randomised order for every repetition. Presentation of the auditory continuum steps was blocked by speaker and the order of presentation for the two speakers was counterbalanced across participants. Auditory input was presented over Sennheiser HD280 headphones at a fixed level. Participants indicated whether they had heard /a:ba/ or /a:da/ by clicking with the computer mouse on labelled buttons on the computer screen. The need for both speed and accuracy was stressed. New trials started after a response was given. The closest step to each participant's 50% cut-off point between /b/ and /d/ for both speakers were selected as their $A_?$ tokens for use in the rest of the experiment.

In the exposure phase, participants viewed the audiovisual tokens $A_?V_b$ and $A_?V_d$ that consisted of natural visual speech tokens combined with the selected auditory step. Participants viewed both the audiovisual /b/ token and the audiovisual /d/ token and presentation of the two tokens was blocked by exposure condition. Blocks were presented in a randomised order. Within each block, the same audiovisual token was presented eight times. There was no explicit task for participants to perform but they were instructed to pay close attention to what the speaker was saying.

An auditory-only test block directly followed each audiovisual exposure block. In the auditory-only test blocks, participants categorised their most ambiguous step ($A_?$) and the two tokens that were one step closer to either end of the continuum ($A_{?-1}$ and $A_{?+1}$). These three auditory tokens were presented twice in a newly randomised order, blocked by repetition. The test tokens were either produced by the same speaker as the participants had heard and seen in the audiovisual exposure task or by a novel speaker. The novel speaker was the other speaker that participants had heard during the auditory-only calibration phase but whom they had not seen in the audiovisual exposure task. Participants categorised the three ambiguous steps as either /a:ba/ or /a:da/ by pressing as quickly and as accurately as possible the button on a button box that corresponded with the respective label shown on the

computer screen ("b" for /a:ba/ and "d" for /a:ba/). In total, participants completed 32 repetitions of an exposure phase followed by a test phase.

2.1.4. Analysis

Results were analysed with linear mixed-effect models in the R statistical package (R Development Core Team, 2007), using the lmer function of the lme4 library (Bates & Sarkar, 2007). The dependent variable was the binomial response to continuum steps (0 = [b]; 1 = [d]). A logistic linking function was used for the categorical dependent variable. The best-fitting model was established by systematic model comparison, using likelihood-ratio tests. We started with a full model and then gradually removed factors that did not contribute to a better model fit, from factors with the largest p values on. Main effects were only removed if their factors did not contribute to an interaction. The best-fitting model included participant as a random factor. Exposure condition (/b/ exposure vs. /d/ exposure) and speaker familiarity (exposure speaker vs. novel speaker) were evaluated as contrast-coded fixed factors. Auditory test token was evaluated as a numerical fixed factor, centred on $A_?$.

2.2. Results and discussion

Participants were sensitive to the fact that the three auditory test tokens formed a continuum, giving more [d] responses to the more /d/-like auditory token than to the more /b/-like token or the most ambiguous token (β = 1.25, SE = 0.05, p < .001; see Figure 2). Overall, participants made more [d] responses to the continuum of the exposure speaker than to that of the novel speaker (β = -0.71, SE = 0.07, p < .001). Participants gave more [d] responses after /d/-exposure blocks than after /b/-exposure blocks (β = 1.46, SE = 0.08, p < .001), indicating that there is an effect of perceptual learning. This effect was found for both the exposure speaker (β = 2.41, SE = 0.12, p < .001) and the novel speaker (β = 0.44, SE = 0.12, p < .001). The size of the

perceptual learning effect is significantly smaller for the novel speaker than for the exposure speaker (β = -2.29, SE = 0.15, p < .001), however, which suggests that learning can generalise across talkers, but that generalisation is not fully realised.

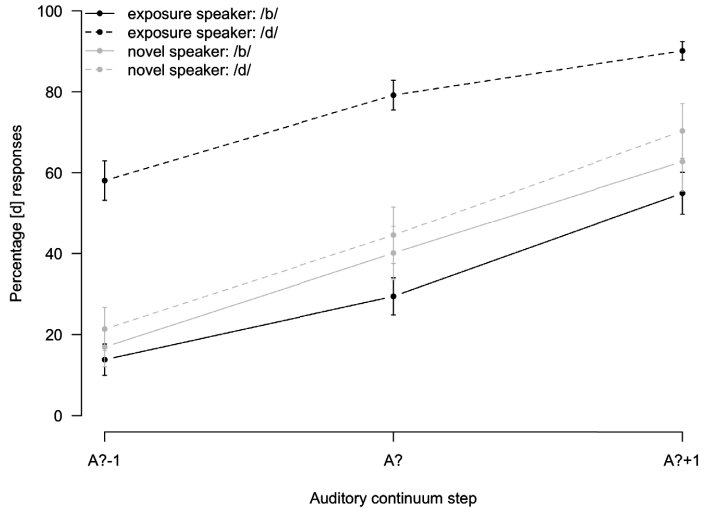

Figure 2: Mean percentages of [d] responses as a function of auditory continuum step in Experiment 1. Solid lines show the results after exposure to $A_?V_b$ and dashed lines after exposure to $A_?V_d$. Black lines show the results for the exposure speaker at test and gray lines for the novel speaker at test.

Visual speech input thus results in the retuning of phonetic category boundaries relative to the learning condition (Bertelson et al., 2003). The results of Experiment 1 indicate that visually guided phonetic retuning affects the identification of speech produced both by the exposure speaker and by a different speaker even when the disambiguating visual speech signal contained information about the identity of the speaker. Generalisation was apparently not fully realised and the effect of retuning is smaller for the novel speaker than for the exposure speaker, which we ascribe to the availability of the speaker information in the visual speech signal.

3. Experiment 2

The results of Experiment 1 indicate that listeners' retuned phonetic category boundaries affected their subsequent identification of speech produced by the exposure speaker as well as by a novel speaker. Generalisation across speakers occurred even when the disambiguating signal contained information about the identity of the speaker. Explicit knowledge about the identity of the speaker thus did not prevent generalisation. The fact that generalisation was not fully realised suggests that the presence of identity information in the visual speech signal may have affected the extent to which transfer occurred. As discussed above, selective adaptation reflects a different change within the perceptual system, namely one due to acoustic input alone and not affected by visual speech information. Recall the results for the McGurk stimuli discussed earlier (Saldaña & Rosenblum, 1994), which revealed that selective adaptation follows the acoustic input even when the perceived utterance differs due to the influence of visual speech information. In Experiment 2, we tested whether the lack of modulation from the visual speech input means that selective adaptation fully generalises to the perception of speech from a novel speaker. As generalisation for selective adaptation has yet to be investigated in previous research, participants in Experiment 2 completed both auditory-only and audiovisual exposure blocks. In the case of a lack of generalisation in the audiovisual exposure condition, the results for the auditory-only exposure condition will help to determine whether this should be ascribed to the presence of visual speech information or to the fact that selective adaptation does not generalise at all. In both the auditory-only and the audiovisual exposure condition, listeners were exposed to fully unambiguous items before completing a categorisation task as in Experiment 1. Whereas the processing of visual speech information is necessary to disambiguate the auditory input in Experiment 1, this is not the case for the unambiguous input in Experiment 2. The effect of retuning is expected to be similar for both the exposure speaker and the novel speaker regardless of the presentation condition in the

exposure phase. This finding would then also provide further evidence for the dissociation between phonetic retuning and selective adaptation.

3.1. Methods

3.1.1. Participants

Twenty-eight new participants (mean age = 21.5; 5 males) from the same population as in Experiment 1 were tested. One participant was excluded due to equipment failure. Another three participants were excluded due to insensitivity to the auditory-only continua. The final data set consisted of the data from 24 participants.

3.1.2. Materials

The materials for Experiment 2 were the same as in Experiment 1. The only difference was that participants were presented with unambiguous auditory-only (A_b and A_d) and audiovisual (A_bV_b and A_dV_d) versions of the exposure materials used in Experiment 1. The audiovisual stimuli consisted of the same unambiguous videos as used in Experiment 1, now combined with the endpoints of the exposure speaker's auditory continuum. The auditory-only exposure materials were created by replacing the video track of the unambiguous audiovisual video tokens with a black frame. Both the audiovisual and the auditory-only exposure stimuli were presented in .avi format. Exposure materials for Experiment 2 were thus entirely free of conflict or ambiguity.

3.1.3. Design and procedure

Experiment 2 differed from Experiment 1 in that participants were presented with unambiguous auditory-only or audiovisual versions of the stimuli during exposure. The experiment again started with an auditory-only categorisation task in which the continua for both speakers were categorised, and $A_?$ was again selected for use in the auditory-only test phase. Following the pretest, participants were exposed

to either the auditory-only A_b and A_d stimuli or to the audiovisual A_bV_b and A_bV_d stimuli. Presentation of the stimuli was blocked by exposure condition; blocks were presented in randomised order. Within each block, the same audiovisual or auditory-only token was presented eight times and participants had no explicit task to perform. They were, however, instructed to pay attention to what the speaker was saying at all times. Every exposure block was immediately followed by an auditory-only test block in which participants performed a categorisation task on $A_{?-1}$, $A_?$ and $A_{?+1}$. Participants completed 32 repetitions of exposure phase followed by test phase.

3.1.4. Analysis

Results were analysed as for Experiment 1. Exposure condition (/b/-exposure material vs. /d/-exposure material), speaker familiarity (exposure speaker vs. novel speaker), and presentation condition of the exposure material (auditory-only vs. audiovisual) were evaluated as contrast-coded fixed factors. Auditory-only continuum step was evaluated as a numerical factor centred on the middle step. Participants were included as a random factor in the best-fitting model.

The analysis of the full model revealed a four-way interaction between the fixed factors (β = 0.87, SE = 0.40, $p < .05$), indicating that the effect of selective adaptation varied as a joint function of talker familiarity, presentation condition, and auditory continuum step. We therefore report the results for the auditory-only test data separately for the auditory-only and the audiovisual exposure conditions.

3.2. Results and Discussion

3.2.1. Auditory-only exposure

Participants gave fewer [d] responses in the auditory-only categorisation test phase after exposure to the auditory-only /d/ token than after the auditory-only /b/ token (β = -1.24, SE = 0.11, $p < .001$; see Figure 3) showing an effect of selective adaptation for the auditory-only materials. Overall, participants made more [d] responses to the continuum of the novel speaker than to that of the exposure speaker

(β = 1.15, SE = 0.11, p < .001). Participants were sensitive to the auditory continua and gave more [d] responses to the more /d/-like test token than to the more /b/-like test token or the most ambiguous step (β = 1.14, SE = 0.07, p < .001). There was a marginally significant difference between the effect of selective adaptation for the exposure speaker and for the novel speaker ($\chi^2(1) = 3.11$, $p = .08$), showing that selective adaptation generalised across speakers with auditory-only exposure materials. There was also a marginally significant difference between the effect of exposure for the three ambiguous test tokens ($\chi^2(1) = 3.75$, $p = .05$), indicating that the shift was larger for the middle step than for the two neighbouring steps.

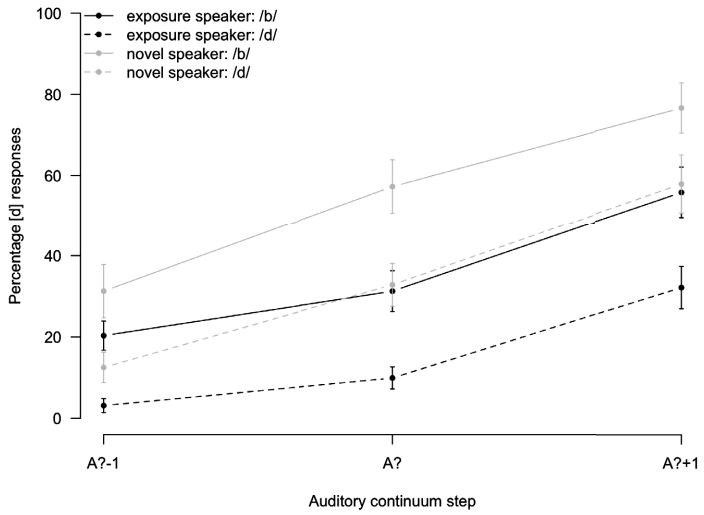

Figure 3: Mean percentages of [d] responses as a function of auditory continuum step following auditory-only exposure in Experiment 2. Solid lines show the results after exposure to A_b and dashed lines after exposure to A_d. Black lines show the results for the exposure speaker at test and gray lines for the novel speaker at test.

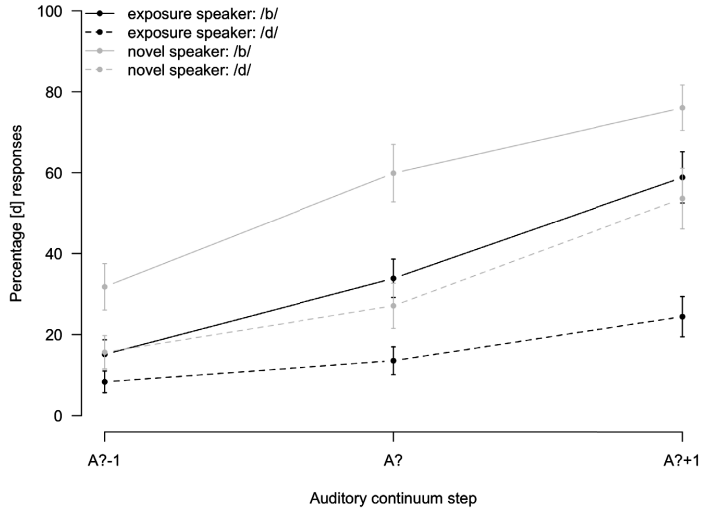

Figure 4: Percentages of [d] responses as a function of auditory continuum step following audiovisual exposure in Experiment 2. Solid lines show the results after exposure to A_bV_b and dashed lines after exposure to A_dV_d. Black lines show the results for the exposure speaker at test and gray lines for the novel speaker at test.

3.2.2. Audiovisual exposure

Participants gave fewer [d] responses in the auditory-only categorisation test phase after exposure to the audiovisual /d/ token than after the audiovisual /b/ token (β = -1.28, SE = 0.11, $p < .001$; see Figure 4), indicating an effect of selective adaptation for the audiovisual materials. Overall, more [d] responses were again given to the novel speaker's than to the exposure speaker's continuum (β = 1.07, SE = 0.10, $p < .001$). Participants were sensitive to the auditory continua giving more [d] responses the more /d/-like the test token (β = 1.02, SE = 0.07, $p < .001$). There was no difference in the selective adaptation effect for the exposure speaker and the novel speaker (not a predictor, $\chi^2(1) = 0.03$, $p = .86$), indicating that selective adaptation generalised across speakers even after exposure to the audiovisual materials. Cross-speaker generalisation of selective adaptation was thus not hindered by the presence of speaker identity information in the visual speech input. There was no difference in

the results for the exposure speaker across the auditory-only and audiovisual exposure conditions (presentation condition not a predictor, $\chi^2(1) = 0.03$, $p = .87$), which provides further evidence for the lack of influence from the visual speech input.

The results of Experiment 2 show that exposure to unambiguous auditory and audiovisual speech made participants less likely to assign ambiguous auditory tokens to the same phonetic category as the phoneme that had been encountered during exposure. In accord with the phonetic retuning results in Experiment 1, we find that selective adaptation affects the interpretation of speech from both the exposure speaker and a novel speaker. The effect of selective adaptation is fully generalised across speakers for both the auditory-only and the audiovisual condition, which is in contrast with the results from Experiment 1 where we observed that the effect of phonetic retuning was smaller for the novel speaker than for the exposure speaker. The availability of visual speaker information during exposure does not appear to affect generalisation of selective adaptation, in line with earlier studies showing selective adaptation to be a purely auditory phenomenon unaffected by visual speech input (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994).

4. General discussion

Adjustments within the perceptual system occur after exposure to speech both when the speech in question is ambiguous and when it is unambiguous. Exposure to ambiguous, idiosyncratic speech results in shifts in listeners' category boundaries in order to incorporate ambiguous sounds into the intended categories. Unambiguous speech input results in decreased sensitivity to particular features of the input when the same sound is presented multiple times. The effects of phonetic retuning and selective adaptation both reflect changes in the perceptual system and indicate that different speech input can have very different results. The results of the current study show that changes in the perceptual system caused by phonetic

retuning and selective adaptation affect the processing of speech from both the exposure speaker and from a novel speaker. Generalisation in both cases occurred despite the availability of speaker identity information in the audiovisual speech input. While selective adaptation fully generalised across speakers in both auditory-only and audiovisual exposure conditions, the generalisation of phonetic retuning was reduced after audiovisual exposure. The availability of speaker identity information may thus have hindered generalisation even though it did not entirely prevent it.

That visually guided retuning of auditory phonetic categories generalises to the identification of speech from a different speaker is in line with previous results for the generalisation of lexically guided retuning (Kraljic & Samuel, 2006). Using lexically guided retuning, generalisation has been shown for plosives (Kraljic & Samuel, 2006) but was not observed for fricatives (Eisner & McQueen, 2005). The apparent discrepancy in these findings was ascribed to the fact that plosives are more invariant across speakers than fricatives and thus provide greater scope for generalisation (Kraljic & Samuel, 2005, 2006). This explanation does not, however, reveal whether it is information about the identity of the speaker or the lack of acoustic similarity that prevented generalisation.

To tease these two alternative explanations apart, our study investigated visually guided retuning rather than lexically guided retuning, and examined a place of articulation contrast for plosives as putatively offering the best chance of cross-speaker generalisation. A voicing contrast could not be used here because the visual speech input was the source of the disambiguating information and listeners are generally unable to distinguish voiced and voiceless sounds on the basis of visual speech (Bernstein et al., 2000; Van Son et al., 1994). The results of Experiment 1 show that the presence of speaker identity information, available in the visual speech input, does not prevent the generalisation of phonetic retuning across speakers. Shifts in listeners' category boundaries affected their subsequent perception of speech even when produced by a novel speaker. Listeners could also not have

disregarded the visual speech input, since it provided the only source of disambiguation for the ambiguous auditory speech input. These results thus suggest that it is acoustic similarity and not the lack of speaker identity information that allows generalisation of retuning across speakers to occur.

Information about the identity of the speaker may, however, have reduced the extent to which generalisation was realised. The effect of retuning on participants' subsequent categorisation of sounds was much smaller, though still statistically significant, for the novel speaker than for the exposure speaker. Such a difference between the results for the exposure speaker and the novel speaker was not observed for the generalisation of lexically guided retuning (Kraljic & Samuel, 2005); but this earlier study did not make explicit whether full generalisation occurred because of the acoustic similarity between the two speakers or due to the lack of information about the identity of the speaker in the input specifically. In the present study, an auditory-only exposure condition combining ambiguous plosive sounds with a non-visual source of speaker identity information could have provided additional information. This was not a possible option, however, given the setup of the current experiment wherein visual speech information was necessary for disambiguation.

Generalisation of phonetic retuning across speakers seems beneficial for listeners since adjusting to speaker's idiosyncrasies brings with it additional costs of processing (Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Nygaard et al., 1994). Speech from every speaker undergoes a process of normalisation using up attentional resources, resulting in recognition being slowed down or becoming less accurate. Listeners can avoid these additional processing costs by applying the changes in the perceptual system established on the basis of speech from one speaker to the identification of speech from another speaker whenever relevant, thereby streamlining the recognition process. Explicit knowledge about the identity of the speakers does not prevent generalisation as long as the auditory input from both speakers is acoustically similar.

Acoustic similarity is also relevant for the generalisation of selective adaptation. Assuming that selective adaptation effects are due to fatigue in the perceptual system, the decrease in sensitivity to phoneme features that characterises the effect only influence the perception of sounds that are acoustically similar to the exposure sound (Eimas & Corbit, 1973). The effect of selective adaptation generalises across phonemes and shows that exposure to /ba/ affects the subsequent perception of both a /ba/-/pa/ continuum and a /da/-/ta/ continuum (Eimas & Corbit, 1973). Generalisation is not observed across position in the syllable, however, which is attributed to the high variability of sounds across syllable positions (Ades, 1974). The results of Experiment 2 show that selective adaptation generalises across speakers when the target sounds were plosives. The auditory-only results suggest that the decrease in sensitivity that occurs after repeated exposure to an unambiguous utterance affects the subsequent perception of speech for both the exposure speaker and a novel speaker. The change in the perceptual system is thus generally applicable and not specific to any speaker.

Selective adaptation also occurs for elements of vision (Webster, 2004; Webster & MacLin, 1999) and recent research has shown that selective adaptation can occur after exposure to static representations of speakers' mouth shapes (Jones et al., 2010). Seeing multiple repetitions of a picture showing a speaker produce a sound thus makes people less likely to perceive a more ambiguous mouth shape as representing that same sound. More interestingly, this effect of selective adaptation to visual speech was the same whether subsequent judgements were given for the exposure speaker or for a novel speaker. This second finding suggests that the selective adaptation effect in this case is not dependent on the identity of the speaker and thus reflects a more general change in the perceptual system. The results of the audiovisual exposure condition in Experiment 2 show a similar effect of generalisation across speakers for selective adaptation to audiovisual speech. Here, too, information about the identity of the speaker was available but did not affect the generalisation of selective adaptation. A decrease in sensitivity to auditory phonetic

features thus affects subsequent perception of these features irrespective of the identity of the speaker.

The generalisation across speakers for selective adaptation was neither prevented nor even reduced by the presence of visual speaker identity information, which is unlike the results for phonetic retuning in Experiment 1. This difference could be due to the fact that in Experiment 2 the visual speech information was not necessary for disambiguation, since the auditory input was unambiguous. However, this finding is also in line with results from earlier studies that have found that selective adaptation is a purely auditory phenomenon and is not modulated by visual speech input (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). These studies used McGurk stimuli and found that the effect of selective adaptation was always in line with the auditory input, regardless of the fact that the perception of the combined audiovisual stimuli was different from the auditory input. The results of the audiovisual exposure condition in Experiment 2 provide further evidence that visual speech information does not affect selective adaptation. Selective adaptation was observed for both the exposure speaker and the novel speaker despite the fact that the visual speech input again contained information about the identity of the speaker.

5. Conclusions

Phonetic retuning and selective adaptation thus affect subsequent recognition of speech produced by the speaker whose speech initiated the changes in the perceptual system. Both effects also influence the recognition of speech from other speakers when the sounds they produced were acoustically similar. Where the two effects diverge is on the extent to which they are generalised to different speakers when information about the identity of the speaker is available. Phonetic retuning generalises across speakers but appears to be hindered somewhat by the presence of speaker identity information in the visual input. On the other hand, this is not the case for selective adaptation and here generalisation occurs to its full extent.

The perceptual system is thus flexible enough to adjust to speech input and does so regardless of whether the input is ambiguous or not. Ambiguous speech input and unambiguous speech input change the perceptual system in different directions, however, as is reflected in the effects of phonetic retuning and selective adaptation. These changes affect how listeners perceive speech on later occasions and this is true for both speech produced by the speaker for whom the original adjustments were made and for speakers who produce acoustically similar sounds. Generalisation across speakers occurs even when listeners have explicit information that the speech they are provided with is produced by a novel speaker. Generalisation for phonetic retuning may be beneficial for listeners as it can reduce processing costs. For selective adaptation, the generalisation indicates that when sensitivity to particular features of a sound is decreased it affects all sounds sharing that feature, whoever produced them. Changes in the perceptual system thus occur for various reasons and show how the system can flexibly adjust to the input it is given.

Chapter 4:

Hearing words helps seeing words:
A cross-modal word repetition effect

Van der Zande, P., Jesse, A., & Cutler, A. (Under revision). Hearing words helps seeing words: A cross-modal word repetition effect. *Speech Communication.*

Abstract

Watching a speaker say words can benefit subsequent auditory recognition of the same words. In this study, we used a cross-modal long-term repetition-priming paradigm to investigate the underlying lexical representations involved in both listening to and seeing speech. We tested whether the auditory presentation of words facilitates their subsequent phonological processing from visual speech. If so, then the two modalities share amodal phonological lexical representations. Additionally, we tested whether speaker repetition influences the magnitude of repetition priming. In Experiment 1, listeners identified auditorily presented words during exposure and visually presented words at test. Test words had occurred during exposure or were new and were produced by the exposure speaker or a novel speaker. Results showed a significant effect of cross-modal repetition priming that was unaffected by speaker changes. In Experiment 2, listeners performed an additional explicit recognition memory task in the test phase. Identification results for Experiment 2 replicated those for Experiment 1. Listeners' lipreading performance can thus be improved by prior exposure to the auditory word forms. Explicit recognition memory, however, was poor, and neither word repetition nor speaker repetition improved it. This suggests that cross-modal repetition priming is not mediated by explicit memory nor is it improved by speaker identity information. Our results indicate that lexical phonological representations are indeed amodal so that they can be shared across auditory and visual processing, but that speaker identity information cannot be transferred at the lexical level.

1. Introduction

Listeners encounter speech produced by a large number of different speakers, who all have their own specific idiosyncrasies due to their particular physiological features (Ladefoged, 1980; Laver & Trudgill, 1979; Mullennix, Pisoni, & Martin, 1989) and their dialectal or sociological backgrounds (Foulkes & Docherty, 2006). Despite this speaker variability, spoken word recognition is generally quick and accurate regardless of the specific surface forms of words. Listeners show improved processing of words that have been previously perceived (Ellis, 1982; Jackson & Morton, 1984; Schacter & Church, 1992) and can benefit especially when words are repeated by the same speaker rather than by a different speaker (Goldinger, 1996; Luce & Lyons, 1998; Mullennix et al., 1989; Schacter & Church, 1992). This indicates that listeners acquire speaker-specific knowledge that then facilitates the subsequent recognition of words produced by the same speaker. In the present study, we examined the effect of spoken word repetition in an auditory-to-visual priming paradigm to investigate whether representations in the mental lexicon are specific to a speech modality, or are amodal and can thus be accessed from both auditory speech and visual speech, and we further investigated the influence of speaker repetition on auditory-to-visual priming to determine whether specific details of a previous utterance are encoded separately, or together with the lexical representations.

Spoken-word recognition can be helped by not only hearing a speaker but also seeing the speaker. Listeners typically benefit in recognising speech when they also obtain such visual speech information (Helfer & Freyman, 2005; Macleod & Summerfield, 1987; Reisberg, McLean, & Goldfield, 1987; Sumby & Pollack, 1954). The benefit of visual speech information is particularly noticeable in situations where the auditory signal is difficult to interpret (Sumby & Pollack, 1954), but information from both sources is always integrated (Arnold & Hill, 2001; McGurk & MacDonald, 1976; Reisberg et al., 1987). Visual speech facilitates the recognition of phonemes and words by providing information that is complementary and redundant to the

auditory signal (Grant, Walden, & Seitz, 1998; Jesse & Massaro, 2010; Summerfield, 1987; Walden, Prosek, & Worthington, 1974). Movements of non-oral facial features (e.g., the eyebrows) and the entire head can further facilitate speech perception by providing prosodic information (Cvejic, Kim, & Davis, 2012; Davis & Kim, 2006; Hadar, Steiner, Grant, & Clifford Rose, 1983, 1984; Krahmer & Swerts, 2004; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). The visual speech signal thus provides the perceiver with an important source of information for spoken word processing.

In order to recognise speech from either an auditory or a visual signal, perceivers access stored representations of words. An important question is whether both speech signals call on the same word representations, or whether the two modalities access separate, modality-specific representations. Perceivers compare an incoming speech signal to lexical representations stored in memory, with lexical items considered as viable candidates for recognition to the degree that they match the signal (Goldinger, 1996, 1998; McClelland & Elman, 1986; Norris & McQueen, 2008). Previous selection of a lexical item facilitates subsequent recognition of the same item (Church & Schacter, 1994; Ellis, 1982; Jackson & Morton, 1984; Tenpenny, 1995). This word repetition priming effect is also observed across modalities, with auditory words being recognised more efficiently when they follow a visual-only presentation of the same word (Buchwald, Winters, & Pisoni, 2009; Kim, Davis, & Krins, 2004). The processing of auditory and visual speech thus appears to call on the same (amodal) representations in the perceiver's mental lexicon. In the present study, we used auditory-only primes followed by visual-only targets to investigate whether auditory and visual speech processing truly rely on amodal lexical representations. We expected to find similar results of cross-modal repetition priming with this paradigm as have been observed with the visual-to-auditory paradigm.

In previous cross-modal repetition priming studies, priming has been short-term (i.e., target immediately following prime). Such studies provide no information about persistence of the repetition-induced facilitation. We used a long-term

(auditory-to-visual) priming paradigm in order to assess whether priming across modalities is long lasting. Long-term word repetition priming effects occur for auditory-to-visual and visual-to-auditory priming when a semantic categorisation task is used (Dodd, Oerlemans, & Robinson, 1989), but that particular task sheds no light on whether the priming is phonological or semantic in nature. Results from short-term visual-to-auditory priming suggest a phonological locus of the effect (much like auditory-only repetition priming; Norris, Butterfield, McQueen, & Cutler, 2006), since visually presented primes limit the range of phonemes perceivers use even in incorrect identifications of auditory targets (Buchwald et al., 2009). In the present study, the task at test was visual-only word identification, allowing us to investigate not only the long-term auditory-to-visual priming effect but also the locus of this effect.

Moreover, we did not restrict our investigation of priming to speech from a single speaker. Perceivers encounter many different speakers, all with their own way of producing sounds. Speaker variability occurs both in auditory and in visual speech, given that visual speech displays the movements of the articulators that underlie the auditory variability (Yehia, Rubin, & Vatikiotis-Bateson, 1998). Variations across speakers have to be taken into account when matching speech to lexical representations. This may involve normalisation, i.e., removal of variability in the surface form before contact is made with the mental lexicon (e.g., Johnson, 2005; Ladefoged & Broadbent, 1957); note that normalisation implies abstract lexical representations of the canonical phonological forms of words, and no consideration of speaker idiosyncrasies at the lexical level (Jackson & Morton, 1984; Luce & Lyons, 1998; Luce & Pisoni, 1998). An alternative account is that lexical representations include detailed information about the surface forms of previous utterances (Church & Schacter, 1994; Goldinger, 1996, 1998). The incoming speech signal is then compared to a large set of previously encountered realisations of words that have been encoded in the lexicon rather than to unitary abstract representations.

Either way, speaker-related variation in the speech signal makes a call on cognitive resources and reduces both speed and accuracy of processing. Both auditory and visual speech are more accurately recognised with a constant speaker than with the speaker varying from trial-to-trial (Creelman, 1957; Mullennix et al., 1989; Yakel, Rosenblum, & Fortier, 2000). Perceivers retain information about speaker idiosyncrasies after exposure to a speaker's voice and this information facilitates the recognition of speech from the same speaker on subsequent occasions (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). Crucially, speaker-specific knowledge acquired from visually presented speech benefits the subsequent recognition of auditory speech from the same speaker, suggesting that information about speaker idiosyncrasies is also modality-independent or amodal (Rosenblum, 2008; Rosenblum, Miller, & Sanchez, 2007). To put the hypothesised amodality of stored speaker knowledge to further test, we investigated here whether auditory exposure to a speaker's voice improves perceivers' subsequent identification of visually presented words from the same speaker. This reverse effect might not be observed if visual speech is special in its influence on auditory speech processing, in that it provides information about the shape and size of the speaker's vocal tract and how the articulators move to produce sounds (Yehia et al., 1998). Being aware of these details may help to process the speaker's voice later, and the details may not be available in auditory speech (though see Fowler, 1986, 1991; Fowler, Brown, Sabadini, & Welhing, 2003; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985). In that case, hearing a speaker might not affect later processing of the speaker's visual speech. If speaker effects do appear in auditory-to-visual priming, however, they would strongly argue for amodal storage of speaker information (Rosenblum, 2008; Rosenblum et al., 2007).

If previous auditory exposure to a speaker's voice indeed positively affects the subsequent processing of visual speech by the same speaker, the question then arises whether same-speaker repetitions produce more priming than different-speaker repetitions. Effects of speaker repetition on implicit memory for words are typically

taken to indicate that speaker-specific information is used to adjust the representations in the mental lexicon (e.g., Goldinger, 1996; although, see Jesse, McQueen, & Page, 2007). This would suggest that same-speaker repetitions would match lexical representations better and hence prime more effectively than different-speaker repetitions. Abstractionist theories, on the other hand, claim that surface details (e.g., speaker idiosyncrasies) are not considered at the lexical level, and such theories would therefore predict no difference in the amount of priming arising from same- versus different-speaker repetitions. We thus also tested whether speaker repetition influenced the magnitude of the word repetition priming effect.

Auditory-only lexical decision and identification in noise is not always affected by speaker repetition. Schacter and Church (1992) initially exposed listeners to clear speech before testing their identification of words presented in noise and failed to find effects of speaker repetition. Goldinger (1996), on the other hand, presented words in noise both during exposure and during test and obtained speaker repetition effects on auditory identification in noise. Thus the presence or absence of effects of speaker repetition may depend on whether or not the first exposure and repetition contexts are similar. An alternative explanation for why some studies have observed speaker repetition effects on implicit memory while others have not is that speaker-specific information may only influence the magnitude of the priming effect when processing is slow, leading to more opportunity for detailed information about previous episodes to be retrieved (McLennan & Luce, 2005; although see Orfanidou, Davis, Ford, & Marslen-Wilson, 2011; Orfanidou, Marslen-Wilson, & Davis, 2006). In our cross-modal priming study, we presented the auditory primes without noise to provide listeners with clear and unambiguous information about the speakers' idiosyncrasies. Given the nature of this cross-modal task, the retrieval situation has to be different from the encoding situation here. Visual-only word recognition is, however, difficult for the average perceiver, and speaker repetition may be helpful. Finding an effect of speaker repetition on the magnitude of cross-modal priming

would strongly suggest that speaker-specific information is stored with amodal lexical representations.

Though the absence of speaker repetition effects in auditory-to-visual priming would argue against speaker-specific details in perceivers' lexicons, such a finding would of course not preclude information about speaker idiosyncrasies being retained elsewhere. Implicit memory (repetition priming) and explicit memory (knowledge of whether a word was presented before) are differently affected by speaker repetition, with explicit memory for auditory words showing a fairly consistent positive effect of speaker repetition (Craik & Kirsner, 1974; Goldinger, 1996; Palmeri, Goldinger, & Pisoni, 1993). Luce and Lyons (1998) found faster explicit memory decisions ("Was this word in the earlier list?") to same-speaker repetitions than to different-speaker repetitions, despite the fact that repetition priming with the same auditory-only materials produced no such differential effect. Audiovisually presented words are also better recognised as old when the voice of the speaker is preserved (Sheffert & Fowler, 1995); listeners in that study were better able to remember the voice in which sounds were produced than the face of the speaker who produced them. Explicit memory may be more susceptible to changes in the surface form than implicit memory, since hearing a word produced by the same speaker a second time can provide additional contextual cues for recognition memory (cf., encoding specificity: Tulving & Thomson, 1973). Perceivers are certainly able to detect whether the same speaker produced auditory-only words and visual-only words (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Munhall & Buchan, 2004). In the present study, we therefore also included an explicit memory task to assess speaker repetition effects in explicit memory across a changed modality.

In summary, the present study investigated whether cross-modal effects of long-term word repetition priming could be obtained using an auditory-to-visual priming paradigm with an identification task. Finding effects of word repetition priming across these modalities would strengthen previous evidence that the processing of auditory and visual speech involves the same lexical representations. We used long-

term priming in order to see whether cross-modal word repetition priming effects persist over large intervals, and we used an identification task to provide evidence relevant to the phonological locus of the priming effect. Additionally, we tested whether speaker repetition effects occur across modalities. Should auditory exposure lead to facilitation of subsequent visual-only identification for the familiar speaker, this would suggest that knowledge about speaker idiosyncrasies is amodal. Further, if speaker repetition affects the magnitude of repetition priming, this would indicate that information about speaker idiosyncrasies is encoded together with the lexical representations in the lexicon. Finally, if cross-modal effects of speaker repetition appear only in explicit memory, this would suggest that speaker-specific information is stored, but separately from lexical representations.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Fifty-three native speakers of Dutch (mean age = 20.8; 10 male) were paid for their participation in Experiment 1. All participants reported normal hearing and normal or corrected-to-normal vision, and none had received prior explicit training in lipreading. Equipment failure caused the loss of data from six participants. The final data set for analysis came from 47 participants, of whom 23 heard Speaker 1 during the exposure phase and 24 heard Speaker 2. Eleven further participants from the same population took part in a pilot experiment (mean age = 21; all female).

2.1.2. Materials

The initial stimulus set consisted of 195 monosyllabic and disyllabic Dutch nouns, all morphologically simple. Words were selected such that the stimulus set included all ten viseme categories distinguished for Dutch (Van Son, Huiskamp, Bosman, & Smoorenburg, 1994). Visemes are sets of speech sounds that are produced with similar external articulatory configurations, and cannot be conclusively

distinguished from visual evidence alone; Dutch viseme categories are shown in Table 1.

One male and one female speaker of Dutch (Speaker 1 and 2, respectively) were recorded using a Sony DCR-HC1000e camera. Both speakers belonged to the same population as the participants and neither speaker had received specific speech training. Recordings were made in front of a neutral background and the speakers were visible from the top of their shoulders to the top of their head. Audio was recorded simultaneously using two stand-alone Sennheiser MKH50 microphones. The speakers produced multiple tokens of all 195 words in isolation and were instructed to avoid list intonation while speaking. One audiovisual token of each word item was selected by the first author for the pilot study. The videos were digitised as uncompressed avi files (720 × 576 pixels) in PAL format. The auditory signal from the same tokens was used for the auditory-only stimuli; sampling rate for the auditory-only materials was 44.1 kHz.

Table 1. *Viseme Categories (Visually Confusable Sets) of Dutch Consonants and Vowels.*

Consonants		Vowels	
Viseme Category	Phonemes	Viseme Category	Phonemes
{p}	/p, b, m/	{i}	/i, ɪ, e, ɛ/
{f}	/f, v, ʋ/	{a}	/ɛɪ, a, ɑ/
{s}	/s, z, ʃ/	{u}	/u, ʏ, ɔ/
{t}	/t, d, n, j, l/	{o}	/ɔʏ, o/
{k}	/k, r, x, ŋ, h/	{au}	/œy, ɔu/

A pilot experiment was conducted, in which 11 participants from the same population as the participants in the main experiment performed a visual-only identification task on all 195 words presented in random order. Participants were

randomly assigned to lipread one of the two experimental speakers and saw the same speaker throughout. Six participants lipread Speaker 1 and five lipread Speaker 2. Participants' task was to identify the word the speaker produced using visual speech information only and to type in their response using the computer keyboard. Before analysing participants' responses, typographical errors were corrected when it could clearly be determined what the intended response had been (e.g., misspellings and switched characters). Participants' original input was left unchanged whenever a typographical error occurred but it could not be unequivocally established what the intended response had been. Homophones were considered as correct responses. Phonetic transcriptions for all responses were added to the dataset using the Celex lexical database for Dutch (Baayen, Piepenbrock, & Van Rijn, 1993). Responses that did not occur in this database were considered incorrect responses but were not excluded from the analyses. Viseme transcriptions, using the Van Son et al. (1994) categories, were added to the dataset on the basis of the phonetic transcriptions.

As a measure of accuracy, we calculated the overlap between the visemes that occurred in the input and the visemes that participants had provided in their response. This measure is less strict than a measure of phoneme overlap or correct word identification, since viseme categories include multiple phonemes and thus multiple responses may be scored as correct (e.g., answering /p/ to a visually presented /b/ would be correct as both are members of the {p} viseme). The viseme overlap score was calculated by counting the number of visemes in the response that also occurred in the input, divided by the larger of the total number of visemes in either the input or the response. The number of overlapping visemes was always divided by the larger of the two totals to ensure that longer responses could not reach 100% correct simply due to exceeding the length of the input. Syllable boundaries were also counted so that participants' overlap score was higher when they provided an answer with the correct number of syllables. For example, if a participant saw the input *lamp* "lamp" and gave the response *lamp*, their viseme

overlap score would be 100%. If the same participant had given the response *lam* "lamb", the viseme overlap score would be 75%. The response *lampen* "lamps" to *lamp* gave a viseme overlap score of 57%, since only four of the seven total characters in the response (i.e., *lam-pen*) overlap with the input visemes. We also recorded the correct word identification scores.

Table 2. *Mean Percentages of Viseme Overlap Scores in the Visual-only Pilot for the Word Sets Created for Experiment 1 and 2 (with Standard Deviations in Parentheses).*

	Set 1	Set 2	Set 3	Set 4
Speaker 1 (M)	59.46 (14.90)	60.65 (12.63)	60.57 (14.92)	64.10 (10.28)
Speaker 2 (F)	62.91 (15.17)	63.86 (10.50)	62.28 (16.01)	64.47 (13.75)

Two independent samples t-test revealed no difference between the two speakers across the 195 pilot words for the correct word identification scores (Speaker 1: M = 7.08%; SD = 13.71%; Speaker 2: M = 8.10%; SD = 14.29%; t(388) = -0.72, p = 0.47) nor for the viseme overlap scores (Speaker 1: M = 57.73%; SD = 13.74%; Speaker 2: M = 59.07%; SD = 14.75%; t(388) = -0.92, p = 0.36). The 120 words that were lipread most accurately for both speakers were selected for use in Experiments 1 and 2 (see Appendix A). Across the selected 120 target words, independent samples t-tests again showed no difference between the two speakers on the correct word identification scores (Speaker 1: M = 11.08%; SD = 16.05%; Speaker 2: M = 12.50%; SD = 16.41%; t(238) = -0.67, p = 0.50) and the viseme overlap scores (Speaker 1: M = 61.20%; SD = 13.27%; Speaker 2: M = 63.38%; SD = 13.86%; t(238) = -1.25, p = 0.21). These 120 words were divided into four word sets that were matched on their visual intelligibility for both speakers (see Table 2) and on average length in syllables. These lists were used to counterbalance the presentation of all words over the four experimental conditions. A 2 × 4 (speaker × word set) analysis of variance using viseme overlap scores as the dependent variable showed no significant main effects for speaker or word set and no significant interaction

between the factors (all F values < 1). The word sets were rotated through the four experimental conditions in the test phases of Experiment 1 and 2 such that all 120 words occurred in all conditions.

2.1.3. Design and procedure

Participants were tested individually in a sound-attenuated booth. The experiment had two phases: an auditory-only exposure phase and a visual-only test phase. Each phase consisted of an identification task. Participants were informed that there would be two separate phases, but were not told about the nature of the task in the second phase of the experiment. In the exposure phase, the task was to identify 60 auditory-only words spoken by a single speaker. These 60 words were taken from two of the four experimental word sets with sets counterbalanced across participants. Half of the participants heard Speaker 1, the other half heard Speaker 2. Words were presented in random order over Sennheiser HD280 headphones at a fixed level. No noise was added to the auditory input. Participants were informed that a real Dutch word would be presented on each trial and that their task was to identify this word by typing in a response using the computer keyboard. Participants were provided with the opportunity to correct their answer before moving on to the next trial. New trials were initiated when the participant pressed the return key to confirm their answer.

In the test phase, participants performed a visual-only identification task on all 120 words from the four word sets. Sixty of these 120 words had previously occurred in the auditory-only exposure phase and the other 60 words were new. In both cases, half the word items were produced by the exposure speaker and the other half were produced by the novel speaker. There were 30 word items in each of the four experimental conditions (i.e., new words/new speaker; new words/old speaker; old words/new speaker; old words/old speaker). Presentation of words and speakers in each condition was counterbalanced across participants. The presentation order of the 120 experimental items was fully randomised. Participants

were again informed that only real Dutch words would be presented and again asked to type their answer using the computer keyboard. New trials started after participants had confirmed their answer by pressing the return key.

2.1.4. Analysis

Participants' responses were checked for typographical errors. Responses were scored for correct word recognition. In addition, viseme overlap scores were calculated for responses given during the test phase using the same procedure as described in Section 2.2. The resulting data set was analysed using linear mixed-effect models in the R statistical package (R Development Core Team, 2007) using the lmer function of the lme4 library (Bates & Sarkar, 2007). The dependent variable was the binomial correct word identification (correct or incorrect). A logistic linking function was used for this categorical dependent variable. Best-fitting models were established through systematic model comparison using likelihood-ratio tests. Factors that did not contribute to a better model fit were removed from the full model, starting from the factor with the highest *p*-value. All best-fitting models included participants as a random factor. Word repetition (old, new), speaker repetition (old, new) and exposure speaker (Speaker 1, Speaker 2) were evaluated as contrast-coded fixed factors.

2.2. Results and discussion

2.2.1. Exposure phase

Participants' auditory-only word identification scores in the exposure phase were high (M = 95.00%; SD = 5.64%). In order to test whether the exposure results differed by exposure speaker, an lmer analysis was conducted that evaluated exposure speaker as a contrast-coded fixed factor and participants as a random factor. The dependent variable was the binomial word recognition score (correct or incorrect). This analysis revealed no significant effect of exposure speaker (β = -0.05,

SE = 0.28, *p* = .83), showing that the means for Speaker 1 (*M* = 95.56%) and Speaker 2 (*M* = 94.51%) did not differ reliably from each other.

2.2.2. Test phase

Participants' visual-only word identification scores in the test phase were, as expected, relatively low (*M* = 15.71%; *SD* = 6.62%). Participants lipread repeated words more accurately than they lipread new words (β = -0.75, *SE* = 0.08, *p* < .001), indicating an overall effect of cross-modal word repetition priming. The effect of speaker repetition varied by exposure speaker (β = 0.84, *SE* = 0.15, *p* < .001) and the results were therefore further analysed separately by exposure speaker (see Table 3).

Table 3. *Mean Percentages of Correct Word Identification in the Experimental Conditions of the Visual-only Identification Task in the Test Phase of Experiment 1 and 2 (with Standard Deviations in Parentheses).*

		New words		Old words	
		New speaker	Old speaker	New speaker	Old speaker
Experiment 1	Speaker 1	12.64 (7.98)	12.92 (7.04)	22.64 (13.19)	25.23 (11.73)
	Speaker 2	10.46 (6.06)	8.06 (5.47)	17.06 (8.10)	16.81 (10.83)
Experiment 2	Speaker 1	13.75 (7.51)	11.25 (6.80)	23.19 (12.06)	21.25 (11.03)
	Speaker 2	8.75 (6.28)	10.00 (7.74)	12.92 (8.06)	18.47 (11.12)

Participants who heard Speaker 1 during the auditory exposure phase were better at lipreading words that were repeated from the auditory-only exposure phase than they were at lipreading new words (β = -0.72, *SE* = 0.11, *p* < .001). This effect was not influenced by changes in the identity of the speaker ($\chi^2(1) = 1.42$, *p* = .23). The old speaker (i.e., here Speaker 1) was lipread better than the new speaker (Speaker 2; β = -0.40, *SE* = 0.11, *p* < .001). Participants who heard Speaker 2 during the auditory exposure phase were also better at lipreading repeated words than new words (β =

-0.79, SE = 0.11, p < .001) and this cross-modal priming effect was again not affected by speaker repetition ($\chi^2(1)$ = 0.32, p = .57). Participants who heard Speaker 2 during the auditory exposure phase lipread the new speaker (Speaker 1) better than the old speaker (Speaker 2; β = 0.44, SE = 0.11, p < .001), explaining the interaction between speaker repetition and exposure speaker in the combined model reported above. Both groups of participants therefore lipread Speaker 1 better than Speaker 2, irrespective of whom they had heard during exposure, and despite the careful matching of word sets on the visual intelligibility of the speakers.

Additional analyses were performed on participants' ability to identify individual visemes in the visual-only test stimuli. Viseme identification was high, as expected (M = 64.11%; SD = 6.67%). Viseme overlap scores also reveal an overall main effect of word repetition (β = -0.14, SE = 0.02, p < .001). Participants' identification of individual visemes was thus improved by word repetition. Again, analyses were split by exposure speaker because the effect of speaker repetition varied as a function of exposure speaker (β = 0.38, SE = 0.05, p < .001). These results revealed the same pattern as observed for the word identification results: Word repetition benefits viseme recognition, regardless of the exposure speaker (Speaker 1: β = -0.12, SE = 0.04, p < .01; Speaker 2: β = -0.16, SE = 0.04, p < .001). Speaker 1 was again generally more intelligible than Speaker 2, thus reversing the effect of speaker repetition (Speaker 1 as exposure speaker: β = -0.16, SE = 0.04, p < .001; Speaker 2 as exposure speaker: β = 0.18, SE = 0.04, p < .001). There was no interaction between word repetition and speaker repetition regardless of exposure speaker (Speaker 1: $\chi^2(1)$ = 1.87, p = .17; Speaker 2: $\chi^2(1)$ = 0.01, p = 0.90). The viseme overlap results thus show a benefit from prior auditory exposure on lipreading visual speech segments: Previously heard speech affects perceivers' visual identification of individual speech segments.

Overall, the results of Experiment 1 revealed long-term, repetition priming across modalities. Participants were better at identifying words and their parts from

visual speech when they had previously heard the words. This cross-modal effect was found regardless of whether words were repeated by the same or a new speaker. Auditory and visual processing of speech utilise the same amodal representations in the mental lexicon and these representations are not updated to contain speaker-specific information.

3. Experiment 2

Experiment 2 compared speaker repetition effects in auditory-to-visual word repetition priming in implicit and explicit memory tasks. The experiment was identical to Experiment 1 except that, at test, participants were first asked to indicate whether the word they perceived visually was a new word or a word repeated from the auditory exposure phase (explicit memory task) before giving their identification response (identification task, reflecting implicit memory).

3.1. Methods

3.1.1. Participants

Fifty-two new participants from the same population as in Experiment 1 (mean age = 20.5; 9 male) took part in return for payment. Four participants' data were lost due to equipment failure. The final analysed data set consisted of data from 48 participants, of whom 24 heard each speaker during exposure.

3.1.2. Materials

The materials were as in Experiment 1.

3.1.3. Design and procedure

The procedure differed from Experiment 1 only in that, during test, participants also performed a recognition memory task on each trial. Participants indicated after each visual-only presentation whether or not they had encountered

the word during the auditory exposure phase, regardless of the identity of the speaker who produced the word; responses were given by pressing one of two buttons corresponding to labels "old" and "new" on the computer screen, with button assignment counterbalanced across participants. Participants had three seconds to respond. After a response had been given, or after the trial timed out, participants were asked to identify the word by typing in their response as in Experiment 1. For the explicit memory task, the instructions stressed the importance of providing an answer as quickly and as accurately as possible.

3.1.4. Analysis

Typographical errors in participants' responses were again corrected, and results analysed using linear mixed-effect models, as described for Experiment 1. For the recognition memory task, the dependent variable was the binomial recognition memory judgement (correct or incorrect). A logistic linking function was used for this categorical dependent variable. The dependent variables for the identification tasks were word identification scores. For the identification task at test, viseme overlap was also analysed. For both identification and recognition memory word repetition (old or new), speaker repetition (old or new), and exposure speaker (Speaker 1 or Speaker 2) were evaluated as contrast-coded fixed factors. Participants were included as a random factor in all best-fitting models.

3.2. Results and discussion

3.2.1. Exposure phase

Participants' auditory-only word identification scores in the exposure phase were high (M = 96.22%; SD = 2.91%). An lmer analysis evaluated exposure speaker as a contrast-coded fixed factor and participants as a random factor, with the binomial word recognition score (correct or incorrect) as the dependent variable. This analysis revealed that the results differed significantly as a function of speaker (β =

1.20, SE = 0.23, p < .001). Although identification approached ceiling for items spoken by each speaker, there was a numerically small but reliable difference between the scores for Speaker 1 (M = 94.24%) and Speaker 2 (M = 98.19%).

3.2.2. Test phase: Recognition memory

Participants' overall correct word recognition was quite low (M = 48.18%; SD = 6.03%) and was similar following both auditory exposure conditions (Speaker 1: M = 48.65%; SD = 5.05%; Speaker 2: M = 47.71%; SD = 6.95%). The complete model for the recognition memory task showed a significant three-way interaction (β = -0.58, SE = 0.21, p < .01), indicating that the results varied as a joint function of word repetition, speaker repetition, and exposure speaker. The results were therefore analysed separately by exposure speaker (see Figure 1).

Participants who heard Speaker 1 in the auditory-only exposure phase showed a marginally significant crossover interaction between the factors word repetition and speaker repetition (β = -0.27, SE = 0.15, p = .07). Neither the main effect of word repetition (β = 0.05, SE = 0.07, p = .48) nor the main effect of speaker repetition (β = 0.00, SE = 0.07, p = .97) reached significance. Participants who heard Speaker 2 during exposure also showed a crossover interaction (β = 0.30, SE = 0.15, p < .05), but the pattern here is the reverse of that for participants who heard Speaker 1. Again, there was no significant main effect of word repetition (β = 0.09, SE = 0.07, p = .23) or speaker repetition (β = 0.08, SE = 0.07, p = .29). The results for both groups together suggest that when participants see Speaker 1 in the visual-only test phase, they are somewhat better at correctly classifying new words as being new than when they see Speaker 2. Overall, the participants' scores were close to chance, however.

An additional analysis was conducted on participants' recognition memory for only those items for which they afterwards provided correct visual-only word identifications. The results showed no main effects of word repetition ($\chi^2(1)$ = 0.01, p = .93), speaker repetition ($\chi^2(1)$ = 1.02, p = .31), or exposure speaker ($\chi^2(1)$ = 0.01, p =

.98), and no interaction reached significance. Thus participants' ability to correctly identify the word in the visual-only speech did not affect their ability to recognise whether the same word was repeated or new.

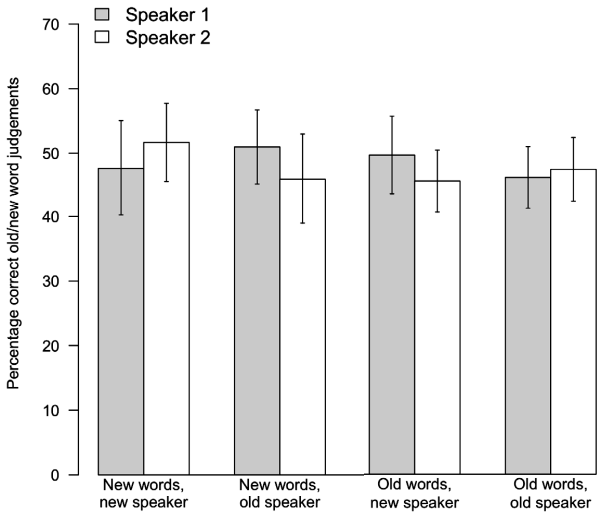

Figure 1: Experiment 2: Mean percentage correct old/new word judgements at the test phase following auditory exposure to Speaker 1 (gray bars) and 2 (white bars) across the four experimental conditions. Error bars show the standard error of the mean.

Participants' sensitivity in the recognition memory task was evaluated by analysing d′ scores, again using linear mixed-effect models. The effect of word repetition could not be evaluated since for the d′ calculations hits were defined as correct "old" responses to old words and false alarms as incorrect "old" responses to new words. The best-fitting model showed no significant main effect of speaker repetition (not a predictor, $\chi^2(1) = 0.38$, $p = .54$) and no significant interaction between speaker repetition and exposure speaker ($\chi^2(1) = 0.24$, $p = .62$). It also showed a non-significant trend of a main effect of exposure speaker ($\beta = -0.27$, $SE = 0.16$, $p = .08$). Participants who had heard Speaker 1 during exposure tended to have better recognition memory performance than those who had heard Speaker 2.

Although the recognition memory results indicated that new visually presented words were more accurately classified as new when spoken by Speaker 1, the d′ results show that participants' ability to recognise whether they had previously heard a word was unaffected by who the speaker was, either at test or during exposure. This finding suggests that the inter-speaker difference in the accuracy data may actually have been due to a bias in responses to the visually presented words.

3.2.3. Test phase: Identification

Participants' visual-only word identification scores in Experiment 2 (M = 14.95%; SD = 7.19%) were approximately at the same performance level as in Experiment 1. The overall results of the visual-only identification task showed a main effect of word repetition (β = -0.67, SE = 0.08, $p < .001$), replicating the cross-modal repetition priming effect of Experiment 1. Participants were better at lipreading words that they had previously heard in the auditory-only exposure phase than words that were new. There was a significant interaction between speaker repetition and exposure speaker (β = 0.81, SE = 0.15, $p < .001$). The results were therefore analysed separately by exposure speaker (see Table 3).

The visual-only identifications for participants who heard Speaker 1 during the exposure phase showed a significant main effect of word repetition (β = -0.64, SE = 0.11, $p < .001$); participants lipread repeated words more accurately than new words. There was also a main effect of speaker repetition (β = -0.49, SE = 0.11, $p < .001$); the repeated speaker was easier to lipread than the new speaker. The word repetition effect was not influenced by speaker repetition ($\chi^2(1) = 2.07$, $p = .15$). Participants who heard Speaker 2 in exposure also lipread repeated words more accurately than new words (β = -0.70, SE = 0.11, $p < .001$) and also showed no significant interaction between word repetition and speaker repetition ($\chi^2(1) = 0.11$, $p = .74$). They showed a main effect of speaker repetition but, as in Experiment 1, the

new speaker (Speaker 1) was lipread more accurately than the old speaker (Speaker 2) ($\beta = 0.33$, $SE = 0.10$, $p < .01$). The speaker repetition effects are apparently driven by differences in visual intelligibility of the two speakers, not by memory factors.

Viseme overlap scores in Experiment 2 were also comparable to those in Experiment 1 ($M = 62.08\%$; $SD = 7.47\%$) and, as expected, higher than the correct word identification scores. Analyses on viseme overlap scores showed a similar pattern of results as the analyses on word scores. There was a main effect of word repetition ($\beta = -0.13$, $SE = 0.03$, $p < .001$) and an interaction between speaker repetition and exposure speaker ($\beta = 0.29$, $SE = 0.05$, $p < .001$). We therefore split the data by exposure speaker and found that participants who had heard Speaker 1 during exposure showed a significant interaction between the factors word repetition and speaker repetition ($\beta = 0.17$, $SE = 0.08$, $p < .05$). This finding indicates that while both the main effect of word repetition ($\beta = -0.13$, $SE = 0.04$, $p < .001$) and the main effect of speaker repetition ($\beta = -0.14$, $SE = 0.04$, $p < .001$) were significant, the advantage of identifying visemes in the repeated words compared to new words was mainly driven by a difference in the old speaker condition. Participants who had heard Speaker 2 during exposure lipread new words better than old words ($\beta = -0.12$, $SE = 0.04$, $p < .01$) and were better at lipreading the new Speaker 1 than the old Speaker 2 ($\beta = 0.29$, $SE = 0.05$, $p < .001$). For these participants there was no significant interaction between the two main effects ($\chi^2(1) = 0.01$, $p = .94$).

The identification results for Experiment 2 largely replicated the results reported for Experiment 1. The main finding is a cross-modal long-term effect of word repetition priming. This repetition priming is observed despite the fact that the repeated words are presented in a different modality on the first and second presentation. The results for the correct word identification are the same across the two exposure groups. For the viseme overlap scores, however, participants who heard Speaker 1 in the exposure task subsequently lipread the visemes in repeated words by the same speaker better than the visemes in repeated words by the novel

speaker. This suggests that in this case speaker repetition enhanced participants' ability to identify individual sounds in cross-modal repetition priming. This same influence of speaker repetition was not observed for participants who had heard Speaker 2 during exposure, however, nor was it observed in the correct word identification data.

4. General discussion

Listeners are able to perceive words more quickly and more accurately when they have been encountered previously (Church & Schacter, 1994; Ellis, 1982; Jackson and Morton, 1984; Schacter & Church, 1992). This facilitation for the processing of repeated words is observed even when there is a change in modality between the first and second presentation of a word (Buchwald et al., 2009; Dodd et al., 1989; Kim et al., 2004). In the present study, we investigated the locus and nature of this cross-modal repetition effect in two experiments by using long-term priming across modalities. In both experiments, listeners first identified words from auditory-only speech and subsequently from visual-only speech. The results show significant repetition priming from auditory speech to visual speech, thus adding to the evidence that both speech modalities share common underlying representations in the lexicon (Buchwald et al., 2009; Kim et al., 2004). Critically, these cross-modal word repetition effects in an identification task suggest that the *lexical* phonological representations are amodal. Moreover, as speaker familiarity did not modulate the size of the effect, the representations must also be abstract.

Experiments 1 and 2 demonstrated that having previously heard a word improved later identification of the same word from visual-only speech. Hearing a word improves the later identification both of the exact word and of the visemes that form that word. The effects of cross-modal word repetition on both word and viseme identification are statistically significant, though numerically relatively small: an improvement of 4-12% for the recognition of the complete word and about 4% for recognition of visemes. The stronger effect is thus on how visemes are interpreted as

a word, suggesting that having heard words before also influences which lexical item is considered the most suitable interpretation of a given the input.

This finding of long-term auditory-to-visual repetition priming extends previous findings of word priming across modalities (Buchwald et al., 2009; Dodd et al., 1989; Kim et al., 2004) in two ways. First, our results show that the identification of words in visual-only speech can also be improved by previous auditory-only exposure. Previous work had focussed on showing a benefit for auditory word recognition after visual-only exposure (Buchwald et al., 2009; Kim et al., 2004). Second, our results provide evidence that long-term auditory-to-visual repetition priming has a phonological locus. Both the visual identification of the segments of a word and the overall identification of the phonological word form benefit from cross-modal word repetition. The processing of visual speech involves the same underlying representations as listeners invoke to process heard speech.

These shared, amodal lexical representations appear to be abstract and to contain no specific details of previously perceived episodes. Auditory word recognition accounts that hold that episodic details about utterances are stored in the lexicon (Goldinger, 1998) predict additional improvement for words repeated by the same speaker over words repeated by a different speaker. While such effects have been observed for auditory-only unimodal repetition priming (Goldinger, 1996), others have failed to find similar effects on auditory-only word recognition (Luce & Lyons, 1998; Schacter & Church, 1992). Our results for cross-modal priming are in line with the latter kind of unimodal studies. Although listeners' visual-only identification performance was better for repeated words than for new words, the magnitude of this effect of word repetition was not modulated by changes in the identity of the speaker. The lack of a speaker repetition effect in cross-modal priming indicates that the underlying representations contacted in identification were sufficiently abstract to allow for variations in the surface details of the repetitions.

An alternative account for this lack of speaker repetition effects, however, might be that speaker information does not transfer across modalities. Speaker-

specific information from auditory-only speech may be encoded in long-term memory without facilitating visual-only speech processing, for instance because indexical information about a speaker is modality-specific even though it is stored with amodal lexical representations. If this is the case, then both speakers perceived during the visual-only identification task in the test phase could be considered new speakers because neither one had previously been perceived visually. Although the transfer of speaker-specific information across modalities has been shown by Rosenblum and colleagues (2007), there are methodological differences between that study and the present one. Most importantly, Rosenblum et al. gave listeners substantially more exposure, in sentences rather than in isolated words. Listeners have been shown to tune in to different speaker-specific properties depending on the kind of speech materials with which they are provided (Cvejic et al., 2012; Grant et al., 1998; Nygaard & Pisoni, 1998), so that speaker-specific information obtained from isolated words could be less susceptible to transfer across modalities than information from sentences. Future research is needed to assess auditory-to-visual transfer of speaker-specific information gained from sentences rather than from words. Also, Rosenblum et al. showed transfer of indexical information from visual to auditory speech, while we examined transfer from auditory to visual speech. It could thus be the case that visual speech, as the source of auditory speech, can provide sufficient information about auditory idiosyncrasies, but auditory speech is not sufficient for defining visual idiosyncrasies. This interpretation is consistent with our own finding that listeners' retuning of auditory phonetic categories by the use of lexical knowledge does not transfer to visual categories unless listeners have also been exposed to a speaker's visual speech (Van der Zande et al., 2013). Such an interpretation of our present results would of course be problematic for theories of speech perception, such as motor theory and direct realism, that suggest that auditory speech input is perceived in terms of the underlying gestures of the speaker's vocal tract (Fowler, 1986, 1991; Fowler et al., 2003; Liberman et al., 1967; Liberman & Mattingly, 1985). If listeners are able to extract such information about

the movements or position of the speaker's articulatory features from the auditory signal, it could be argued that prior experience with a speaker's voice should also benefit subsequent processing of visual-only speech. This was, however, not the case here.

Another alternative explanation is that speaker-specific information can be transferred cross-modally, but this transfer takes place at a prelexical level. Exposure to auditory speaker-specific information has been shown to trigger adjustments of phonetic categories at a prelexical stage of processing (McQueen et al., 2006). Both repetitions of complete words and repetitions of their individual phonemes in unimodal auditory presentation result in facilitation of the later processing of (auditory) speech, albeit to different extents (Jesse et al., 2007). Exposure to speaker-specific information about particular phonemes can thus benefit the subsequent processing of words that are different but contain (some of) the same phonemes. In our study, the new words spoken by the exposure speaker during test contained phonemes that the perceivers had already heard spoken by the same speaker during exposure. That is, on the phoneme level, the new words were (partially) old, since they contained sounds to which perceivers had previously been exposed from the same speaker. If indexical knowledge is applied and transferred across modalities at the prelexical level, then we should have found a main effect of speaker repetition here. The lack of such an effect casts doubt on the suggestion that speaker knowledge is transferred across modalities at a prelexical level. Again this is in line with our lexically guided retuning results (Van der Zande et al., 2013) in which listeners given auditory-only exposure used lexical knowledge to retune auditory phonetic categories, but not the corresponding visual phonetic categories. Adjustment of phonetic categories hence did not transfer across modalities.

Although we found no speaker repetition effect on word repetition priming, we observed an overall speaker effect in our data. The speaker effect is a global benefit for processing spoken words from the visual speech of Speaker 1 over that of Speaker 2, independent of whom the exposure speaker was. Speakers differ in their

intelligibility, in that some speakers may inherently be easier to understand than others (Bond & Moore, 1994; Gagné et al., 1994). It seems unlikely to be the case in our results, however, that visual-only perception for Speaker 1 was inherently easier than visual-only perception for Speaker 2. The 120 word stimuli and the four word sets in which these stimuli were divided for Experiment 1 and 2 were closely matched on visual intelligibility of the speakers and showed no significant difference across the two speakers in the pilot study (see Section 2.2); rather, it was Speaker 2 who was slightly more easy to lipread in the pilot. We conducted additional analyses comparing the results from the pilot experiment with the results from Experiment 1 and 2, specifically focussing on the results from the new words/new speaker condition. These results are similar to those obtained in the pilot experiment, where participants had no prior exposure to either the speaker or the words they had to lipread. Independent samples t-tests using viseme overlap as the dependent variable showed that there was a marginally significant difference between speakers for the 120 word items in the new words/new speaker results in Experiment 1 ($t_1(45) = 1.76$, $p = 0.08$; $t_2(236) = 1.88$, $p = 0.06$). The difference between the two speakers in Experiment 2 was also significant ($t_1(46) = 2.33$, $p = 0.05$; $t_2(236) = 2.72$, $p < 0.01$). In both cases, the viseme overlap scores for Speaker 1 exceeded those for Speaker 2. In the pilot experiment, however, there was no such difference between the results for the two speakers when analysing the results for these 120 word items used in Experiment 1 and 2 ($t_1(9) = -0.34$, $p = 0.74$; $t_2(236) = -1.19$, $p = 0.24$). Analyses comparing the results of the pilot with the results from Experiment 1 and 2 revealed no significant differences, however (all *p*-values > 0.05), confirming that performance in the new words/new speaker condition in Experiments 1 and 2 was similar to that in the pilot for these words. Together with the result that the performance for Speaker 2 was numerically higher in the pilot than for Speaker 1, this seems to make it unlikely that Speaker 1 was inherently easier for participants to lipread.

One further difference between the pilot study and our Experiments is that all 11 pilot study participants were women. This is not highly likely to have affected the

generalisability of the pilot results, especially given that most Experiment 1 and 2 participants were also female (of 105 participants in all, only 19 were male). Male-female differences in audiovisual speech recognition abilities have appeared in some studies (e.g., Strelnikov et al., 2009) but not in others (e.g., Irwin et al., 2006). However, we examined the data for male and female participants separately, and found parallel speaker intelligibility differences for both groups. In Experiment 1, male participants' visual-only identifications for Speaker 1 (male) were better than for Speaker 2 (female), regardless of whether their initial exposure was to Speaker 1 (Speaker 1: M = 64.70%; Speaker 2: M = 59.42%) or to Speaker 2 (58.58% to 50.89%). Female participants showed the same pattern: exposure to Speaker 1, 68.26% to 62.92%, exposure to Speaker 2, 67.13% to 62.26%). In Experiment 2, again, male participants lipread Speaker 1 better than Speaker 2 after either exposure: to Speaker 1, 62.21% to 58.95%, to Speaker 2, 67.40% to 59.35%); female participants did the same: exposure to Speaker 1, 63.46% to 59.29%, exposure to Speaker 2, 64.54% to 60.93%. Our finding that a male speaker (Speaker 1) was easier to lipread than a female speaker (Speaker 2) also does not agree with reports that participants generally find female speakers easier to lipread than male speakers, regardless of the participants' sex (Daly et al., 1997).

Further, differences between visual-only identification scores for Speaker 1 and 2 also were not due to differences in auditory identification of the speaker's speech during the auditory-only exposure. Although we found a significant difference in identification scores for Speaker 1 and 2 in Experiment 2, listeners' auditory performance was worse for Speaker 1 than for Speaker 2. We therefore can only suggest that Speaker 1's advantage in the experimental situation reflects some as yet unidentified dimension of visual articulation that can play a role in recognising articulated versions of previously heard words. This topic deserves further empirical investigation, but does not affect the conclusions drawn from the present study.

The results of Experiment 2 also showed no effects of speaker repetition on explicit recognition memory. Listeners were equally likely to correctly classify words as being old regardless of whether these repeated words were produced by the same speaker in both instances or by a different speaker. This is in contrast to previous findings of a same-speaker advantage in auditory explicit memory studies (Goldinger, 1996). Therefore, it does not seem that speaker repetition improves the explicit memory of repeated words across modalities. Remembering 60 individual words from auditory speech without an explicit prompt to do so may have been a difficult task for participants, although other long-term recognition memory studies, some with even higher numbers of items, have shown that this is not beyond the capability of an average listener (Bradlow et al., 1999; Craik & Kirsner, 1974; Schacter & Church, 1992; Sheffert, 1998). Our participants showed rather poor performance and did not detect word repetitions reliably. We observed, however, two marginally significant interactions indicative of old/new distinctions being more accurate for Speaker 1 than for Speaker 2. This difference may be related to the observed difference in word identification for the two speakers. When visual speech information is more difficult to process recognition memory decisions may also become harder. Alternatively, the absence of speaker effects in the d' analyses also suggests a role for response bias in this difference between speakers. The finding that explicit memory of repeated words across modalities was not facilitated by speaker repetition suggests that the memory for the previously perceived utterances contained no speaker details; however, it may also be the case that listeners could not extract enough information from the visual-only words to trigger explicit recognition memory (though they extracted enough to induce repetition priming).

The perception of just visual speech without auditory information plays only a limited role in our normal interaction with others. Auditory-only communication (e.g., telephone conversation) and audiovisual communication (e.g., face-to-face interaction) are far more likely to occur. Although we may see many people speaking together from afar without ever hearing them, choosing to communicate with

someone through visual-only speech production is much rarer and is most likely with speakers with whom we are familiar and whom we have heard speak before. In most cases, then, visual-only exposure before auditory-only exposure, as investigated by Rosenblum et al. (2007), is unlikely because familiarity with a speaker through auditory speech will usually precede familiarity with a speaker on the basis of visual-only speech. When someone mouths something to us across a busy conference room it would be beneficial for our visual-only identification performance if we could be primed by auditory words perceived earlier. Our results show that such priming across modalities does indeed take place, though it is limited in its extent. In the same situation, our visual-only identification of speech from an *unfamiliar* speaker will also benefit from containing words that we have recently perceived auditorily, showing that when necessary we can even lipread people that we have not heard before.

5. Conclusions

This study investigated the effects of word repetition and speaker repetition on implicit and explicit memory in an auditory-to-visual, long-term cross-modal priming paradigm. The results indicate that both auditory processing and visual processing share lexical representations, because the processing of repeated words is facilitated across speech modalities. These amodal lexical representations are abstract and are not adjusted on the basis of speaker-specific information. Repeated words and their segments are consistently identified better than new words, regardless of the identity of the speaker. Neither implicit memory nor explicit memory of repeated words was enhanced by repetitions being produced by the same speaker on both instances. Speaker-specific information therefore does not appear to be transferrable across modalities at the lexical level.

Appendix

Viseme Overlap for All 120 Experimental Words.

			Pilot		Experiment 1		Experiment 2	
Dutch word	**Dutch transcription**	**English gloss**	**Speaker 1**	**Speaker 2**	**Speaker 1**	**Speaker 2**	**Speaker 1**	**Speaker 2**
baby	ˈbeː-bi	baby	62.22	75.56	63.64	63.23	53.15	66.96
bar	bɑʀ	bar	79.44	57.38	86.44	60.80	84.78	45.84
beek	beːk	brook	50.00	53.33	55.72	81.94	51.32	76.39
bende	ˈbɛn-də	gang	62.04	78.33	45.34	70.19	50.20	67.99
bezem	ˈbeː-zəm	broom	51.59	48.10	55.57	71.70	49.26	78.49
bijbel	ˈbɛɪ-bəl	bible	63.89	59.17	84.00	73.78	72.00	73.94
boef	buf	thug	76.67	40.00	73.96	33.86	67.01	44.10
boel	bul	bunch	58.89	66.67	75.05	68.75	64.51	65.56
bom	bɔm	bomb	78.33	79.50	95.83	51.42	98.61	61.20
bon	bɔn	ticket	67.78	95.00	65.53	81.12	69.93	73.68
boog	boːx	arch	57.22	70.00	76.27	74.38	85.07	64.58
boom	boːm	tree	82.14	88.00	97.71	64.36	92.33	74.70
burger	ˈbʏr-xər	citizen	42.86	42.86	34.42	50.80	37.27	45.13
cheque	ʃɛk	check	58.33	59.00	68.12	70.05	63.26	63.31
chic	ʃik	chic	66.67	46.33	84.91	69.57	77.08	65.69
faam	faːm	fame	60.56	55.00	66.30	52.60	74.03	48.89
fabel	ˈfaː-bəl	fable	78.57	71.43	90.97	64.91	87.70	69.54
fan	fɛn	fan	45.00	63.33	48.12	68.84	55.83	69.73

feit	fɛɪt	fact	85.83	61.33	81.16	49.63	77.78	56.80
fik	fɪk	fire	72.22	85.33	83.33	67.02	76.88	72.71
folder	ˈfɔl-dər	leaflet	59.82	77.14	59.86	49.72	56.55	43.15
fout	fɔut	error	49.35	62.17	65.14	66.39	78.87	65.32
fuif	fœyf	party	62.22	50.67	73.67	45.50	63.67	46.75
fut	fʏt	energy	48.61	55.00	64.47	44.66	51.39	48.06
gaas	xaːs	gauze	52.50	38.67	67.54	39.01	62.92	42.22
gang	xɑŋ	hallway	55.33	78.67	63.12	69.17	63.54	63.61
gevel	ˈxeː-vəl	façade	67.86	71.43	68.45	76.29	66.77	69.74
gif	xɪf	poison	51.94	70.33	68.75	53.13	66.67	56.60
gil	xɪl	yell	41.52	72.50	40.33	55.61	40.36	55.63
gordel	ˈxɔr-dəl	seatbelt	33.33	48.57	48.21	53.42	41.67	42.26
kamer	ˈkaː-mər	room	58.33	63.33	64.49	70.04	59.92	53.99
kater	ˈkaː-tər	hangover	66.67	67.50	67.99	62.67	65.97	53.67
keizer	ˈkɛɪ-zər	emperor	68.25	54.76	48.41	54.61	47.22	52.18
kip	kɪp	chicken	70.56	75.71	67.39	72.35	68.24	82.15
koffer	ˈkɔ-fər	suitcase	62.70	63.33	61.83	57.14	60.88	72.42
kom	kɔm	bowl	55.00	83.33	54.87	62.48	58.28	61.12
kop	kɔp	mug	83.61	66.67	86.81	50.74	84.17	59.08
kuif	kœyf	quiff	26.67	46.67	47.92	45.98	46.81	60.87
lach	lɑx	smile	83.33	86.67	72.22	65.58	70.35	67.69
leger	ˈleː-xər	army	80.56	66.67	74.74	48.61	65.28	52.78

leraar	ˈleː-raːr	teacher	69.44	66.67	75.78	61.56	60.88	63.10
les	lɛs	lesson	65.56	41.33	62.33	59.00	52.14	52.05
liefde	ˈliv-də	love	59.26	63.33	70.14	78.26	56.65	69.25
maag	maːx	stomach	72.22	88.33	79.93	65.58	70.97	62.50
merel	ˈmeː-rəl	blackbird	56.75	80.00	63.10	72.59	61.61	66.91
moed	mut	courage	64.58	61.67	63.57	61.30	61.96	62.29
mok	mɔk	mug	62.78	74.67	62.77	70.36	65.72	72.92
motor	ˈmoː-tər	motorcycle	56.75	73.33	73.51	72.49	67.60	61.31
mug	mʏx	mosquito	65.28	43.33	71.53	75.00	67.85	63.74
muis	mœys	mouse	28.61	58.33	41.81	64.33	37.29	56.71
muur	myr	wall	78.06	88.33	61.41	72.78	69.62	57.38
naad	naːt	seam	59.44	33.05	70.80	44.79	67.15	40.14
nagel	ˈnaː-xəl	nail	61.90	57.50	61.59	50.00	68.45	55.85
neef	neːf	cousin (M)	68.06	39.83	79.86	64.54	77.85	53.45
negen	ˈneː-xə	nine	65.24	72.00	66.19	69.78	63.91	61.07
nek	nɛk	neck	63.89	60.33	57.61	62.85	56.67	67.15
nier	nir	kidney	59.44	73.33	57.15	84.72	56.79	74.16
nis	nɪs	niche	61.11	72.33	77.29	58.49	56.46	59.22
noot	noːt	nut	47.33	86.67	52.46	73.96	50.70	68.54
nummer	ˈnʏ-mər	number	63.89	63.33	73.44	54.37	53.27	60.16
parel	ˈpaː-rəl	pearl	41.67	70.00	58.12	71.74	46.51	52.26
pas	pɑs	pass	65.28	86.67	70.20	79.84	70.27	69.27

paus	pɔus	pope	80.00	63.33	71.01	69.38	66.81	64.18
piek	pik	peak	80.56	81.67	75.16	66.23	71.16	64.77
pijp	pɛɪp	pipe	53.10	74.17	75.43	48.61	72.33	50.95
pit	pɪt	stone	64.09	57.57	77.18	65.86	79.65	67.57
poging	ˈpoː-xɪŋ	attempt	59.23	65.83	52.38	58.34	53.37	54.64
pool	poːl	pole	42.78	78.67	71.88	72.59	75.35	70.42
put	pʏt	well	55.56	68.00	58.33	62.20	56.88	60.91
raad	raːt	advice	55.56	57.33	72.78	47.39	69.20	45.31
reep	reːp	strip	41.67	71.90	51.25	72.28	66.36	81.04
regen	ˈreː-xə	rain	56.67	52.67	59.86	43.47	52.71	47.28
rel	rɛl	riot	50.56	70.00	40.83	58.17	36.81	50.40
riem	rim	belt	72.22	76.90	61.81	63.76	77.08	73.19
ring	rɪŋ	ring	47.22	72.00	50.35	64.49	56.25	54.31
rook	roːk	smoke	31.03	73.67	49.54	55.42	50.14	53.62
rubber	ˈrʏ-bər	rubber	67.17	56.67	74.90	38.61	63.89	54.46
rug	rʏx	back	42.78	55.00	70.83	60.20	68.06	65.56
ruzie	ˈry-zi	row	42.86	57.33	67.47	47.06	60.14	45.27
sap	sɑp	juice	69.44	53.33	68.42	46.91	63.54	55.75
satan	ˈsaː-tɑn	satan	67.46	50.95	60.23	52.03	57.00	54.80
saus	sɔus	sauce	34.72	40.00	41.10	69.24	36.23	56.81
sein	sɛɪn	sign	61.11	49.17	66.01	47.69	67.71	58.48
set	sɛt	set	75.48	50.79	74.17	57.27	63.02	47.30

shampoo	ˈʃɑm-poː	shampoo	51.39	45.24	78.34	51.43	81.25	47.45
sik	sɪk	goatee	66.67	31.67	79.62	65.30	72.83	68.47
sinas	ˈsi-nɑs	orange soda	58.33	54.29	56.14	65.36	51.12	62.24
soep	sup	soup	70.00	55.00	74.64	75.28	64.58	69.38
suiker	ˈsœy-kər	sugar	51.19	66.67	46.14	74.68	38.42	60.76
taak	taːk	task	73.61	52.00	77.90	48.97	79.58	48.00
taal	taːl	language	38.89	43.33	36.11	28.24	47.15	30.95
tafel	ˈtaː-fəl	table	97.22	67.14	95.96	71.63	92.66	71.97
tak	tɑk	branch	56.67	68.33	76.04	65.49	82.64	61.31
tempel	ˈtɛm-pəl	temple	78.87	73.77	77.10	68.06	73.81	66.15
titel	ˈti-təl	title	50.00	42.22	54.96	61.73	57.74	60.86
toeval	ˈtu-vɑl	coincidence	63.89	66.19	71.22	42.34	60.91	59.82
toon	toːn	tone	51.39	45.33	57.99	42.33	49.93	41.62
vaas	vaːs	vase	70.56	42.86	72.60	59.28	81.80	45.66
vak	vɑk	square	70.56	61.67	79.24	68.35	74.64	54.87
val	vɑl	fall	38.06	32.33	45.14	44.71	57.99	37.64
vat	vɑt	barrel	71.11	85.00	64.44	46.35	66.83	54.58
veer	veːr	feather	66.67	83.33	70.83	73.20	69.72	68.25
vijf	vɛɪf	five	77.78	90.00	85.42	79.30	70.49	77.08
voedsel	ˈvut-səl	food	73.02	68.57	69.94	63.74	59.33	66.07
voeg	vux	joint	51.51	57.90	82.23	74.58	68.91	78.59
vogel	ˈvoː-xəl	bird	78.57	60.12	92.64	82.61	93.48	77.98

vuur	vyr	fire	51.79	56.29	73.08	66.72	72.00	50.84
zaad	zaːt	seed	70.00	55.00	62.63	46.27	61.76	46.05
zak	zɑk	sack	77.78	67.00	72.27	66.84	80.43	62.13
zebra	ˈzeː-braː	zebra	45.24	70.48	61.41	61.34	57.77	66.05
zeef	zeːf	sieve	54.37	49.90	64.20	47.45	76.13	55.73
zeep	zeːp	soap	76.39	68.00	78.47	58.85	63.06	70.67
zeil	zɛɪl	tarpaulin	38.33	51.67	60.97	54.48	52.37	50.86
zes	zɛs	six	59.52	75.33	63.29	66.11	63.47	48.13
zet	zɛt	move	65.00	70.50	73.60	59.91	69.06	56.16
zoemer	ˈzu-mər	buzzer	61.90	66.67	61.37	69.18	72.72	60.68
zoen	zun	kiss	65.00	63.33	71.56	62.92	64.58	57.44
zomer	ˈzoː-mər	summer	74.21	70.00	53.66	64.88	59.61	63.91
zuivel	ˈzœy-vəl	dairy	67.86	67.62	54.35	65.87	60.91	67.06
zuurkool	ˈzyr-koːl	sauerkraut	64.29	54.29	59.33	58.94	49.49	60.42

Note. All viseme overlap scores are listed in percentages. The means for Experiment 1 and 2 are calculated over all four experimental conditions.

Chapter 5:

Summary and conclusions

1. Summary

Variation in the way sounds are realised by speakers that we communicate with on a daily basis is ubiquitous. Exposure to the speech produced by these speakers leads to adaptation of the perceptual system of the listener (Bertelson, Vroomen, & De Gelder, 2003; Diehl, 1975; Eimas & Corbit, 1973; Norris, McQueen, & Cutler, 2003). These perceptual adaptations serve as a mechanism enabling us to adjust to differences in speech resulting, for instance, from physiological, sociological, and dialectal backgrounds (Fant, 1973; Foulkes & Docherty, 2006; Laver & Trudgill, 1979; Peterson & Barney, 1952) and such adjustments generally facilitate the recognition of speech (Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). Changes in the perceptual system of the listener occur at various levels of processing and often occur after very little exposure (Norris et al., 2003; Vroomen, Van Linden, De Gelder, & Bertelson, 2007). In the experiments discussed in this thesis, we investigated how perceptual adjustments influence the subsequent processing of auditory and visual speech. Combined auditory and visual speech input constitutes a significant portion of the speech with which listeners are confronted on a daily basis. The consideration of audiovisual speech is hence necessary for a full understanding of how listeners process speech. Audiovisual speech also provides the most complete source of speech information (Massaro & Cohen, 1983; Massaro & Friedman, 1990; Reisberg, McLean, & Goldfield, 1987; Rosenblum, 2005, 2008) and as such may be particularly informative with respect to speaker-specific information. The goal of this thesis was to provide new knowledge about the adjustments that occur in listeners' perceptual system after exposure to auditory and visual speech. Further, we investigated the nature of the information about perceived speech and speakers that is stored by listeners in long-term memory, and how availability of this information affects the subsequent processing of speech.

The first two experimental chapters of this thesis (Chapters 2 and 3) focused on the retuning of phonetic categories. Phonetic categories exist for auditory speech

and visual speech and these categories are used to analyse the incoming speech at a prelexical level of processing. The native language largely shapes listeners' phonetic categories (Kuhl et al., 2006; Narayan, Werker, & Beddor, 2010; Werker & Tees, 1984), although the boundaries between the categories are flexible (Bertelson et al., 2003; Maye, Aslin, & Tanenhaus, 2008; Norris et al., 2003). Exposure to speaker-specific idiosyncrasies can shift (i.e., retune) the boundaries between phonetic categories. An auditory speech signal that is ambiguous between two categories in a speaker's idiolect can be disambiguated by additional information, such as visual speech and lexical knowledge, which then results in the subsequent retuning of the category boundaries (Bertelson et al., 2003; Norris et al., 2003). Phonetic categories are adjusted such that the previously ambiguous idiosyncrasy can now be assigned to the correct phonetic category. Visual phonetic categories can also be retuned. Simultaneously presented auditory speech can function as the disambiguating source that guides the retuning of visual phonetic categories (Baart & Vroomen, 2010). But are such phonetic categories retuned independently for both modalities or can information that shifts boundaries in one modality also lead to shifts in the boundaries between categories for the other modality?

The experiments in Chapter 2 investigated whether lexical information could be used to retune the visual phonetic categories. Additionally, in Chapter 2 we also directly investigated the link between the phonetic categories in the two speech modalities by testing whether the lexically guided retuning of auditory phonetic categories could result in shifts in the boundaries between visual phonetic categories. Evidence for this cross-modal transfer of retuning as guided by lexical knowledge would show that the phonetic categories for both modalities are indeed linked. In order to establish whether lexical knowledge could guide the retuning of the boundaries between visual phonetic categories, listeners in Experiment 2.1 were exposed to Dutch words that contained a phoneme that was auditorily and visually ambiguous. Despite the ambiguity in both modalities, the words in which these ambiguities occurred disambiguated the speech input and no other disambiguating

information was available. Listeners were subsequently tested on their interpretation of visual-only speech containing a similar idiosyncrasy. The results of Experiment 2.1 showed that visual phonetic category boundaries could indeed be retuned by lexical information. Listeners' interpreted more visually ambiguous phonemes as belonging to the phonetic category that was favoured by the lexical context in which these ambiguities had previously been presented during the exposure phase. In Experiment 2.2, we tested whether lexical information can retune visual categories through a retuning of the auditory categories. In other words, can speaker-specific knowledge be generalised across modalities? In Experiment 2.2, listeners were exposed to auditory-only words again containing an idiosyncratic, ambiguous phoneme that was disambiguated by the lexical context and were subsequently tested on their interpretation of either auditory-only or visual-only speech. The results of the auditory-only test phase of Experiment 2.2 indicated that lexical information leads to retuning of auditory phonetic boundaries, thereby replicating results presented in early studies (Norris et al., 2003). The results of the visual-only test phase, on the other hand, showed no influence of lexically guided retuning. Exposure to an auditory ambiguous phoneme resulted in the lexically guided retuning of auditory phonetic categories, but such exposure was thus not sufficient to cause shifts in the boundaries between visual phonetic categories. Visual phonetic categories therefore only showed influences of lexical information when the visual-only ambiguity and the disambiguating lexical information were presented simultaneously during exposure. These findings indicate that phonetic categories for the two modalities of speech are not inextricably linked, and that changes in the phonetic categories for one modality do not automatically result in similar changes in the phonetic categories of the other modality.

The results of Chapter 2 indicated that the phonetic categories for auditory speech and visual speech are separate and that changes in the category boundaries for auditory speech do not result in changes in visual phonetic categories. For the retuning of category boundaries to occur, an idiosyncrasy has to be presented

together with the disambiguating information. Phonetic retuning thus does not generalise across modalities. Previous research has shown, however, that phonetic retuning can generalise across speakers in some cases, depending on the nature of the phoneme contrast (Eisner & McQueen, 2005; Kraljic & Samuel, 2006). Generalisation across speakers did not occur when retuning affected category boundaries for phonemes that vary substantially between speakers, but did occur for sounds that varied across a single dimension (e.g., duration) and thus showed less variation between speakers (Kraljic & Samuel, 2005, 2006). In Chapter 3, it was investigated whether phonetic retuning could also generalise across speakers after audiovisual exposure. This was done so that we could tease apart whether it was acoustic similarity or speaker identity in the exposure materials that allowed for generalisation in the findings reported by Kraljic and Samuel (2006). In Experiment 3.1, listeners were exposed to audiovisual syllables that contained a sound that was auditorily ambiguous but not visually. Listeners categorised auditory-only sounds in the subsequent test phase. The question was whether generalisation across speakers could occur for visually guided retuning of auditory phonetic categories. Again, the visual speech that served as the source of disambiguation contained clear information about the identity of the speaker, in order to disentangle whether it is acoustic similarity or speaker identity information that affects generalisation. The results of Experiment 3.1 showed that the lexically guided retuning of auditory categories affected the processing of speech from both the exposure speaker and the novel speaker. The retuning thus generalised across speakers despite the availability of speaker identity information in the visual speech signal during exposure. This finding suggests that acoustic similarity across speakers predicts whether generalisation of speaker-specific knowledge can occur. The size of the retuning effect was diminished for the processing of speech by the novel speaker, however, which indicated that identity information might have affected generalisation despite not fully preventing it.

In Chapter 3, we also investigated whether selective adaptation to a speaker would generalise across speakers after auditory and audiovisual exposure. Unlike phonetic retuning, which occurs after exposure to ambiguous idiosyncrasies, selective adaptation occurs after repeated presentation of unambiguous sounds (Diehl, 1975; Eimas & Corbit, 1973; Sawusch & Jusczyk, 1981). Selective adaptation, therefore, does not reflect changes in the perceptual system occurring in order to overcome variations in the speech input. But rather, it has generally been assumed that the effects of selective adaptation are due to fatigue in the perceptual system after prolonged exposure to the same acoustic features (Samuel, 1986). Due to this fatigue, listeners become less sensitive to the particular features they have been exposed to and this decrease in sensitivity results in fewer sounds being interpreted as being part of the category from which the exposure sounds were drawn. In Experiment 3.2, listeners were exposed to unambiguous auditory and audiovisual syllables. Auditory speech was used in order to establish whether selective adaptation that followed exposure to one speaker could also influence the interpretation of the other speaker in unimodal speech recognition. We again tested whether speaker identity information in the input could affect the generalisation of this effect across speakers by using audiovisual speech during exposure. The results of Experiment 3.2 showed generalisation of selective adaptation across speakers after auditory-only exposure and after audiovisual exposure. Again both the interpretation of speech by the exposure speaker and speech by the novel speaker were affected by the adjustments that occurred in the perceptual system after exposure. Unlike phonetic retuning, the results for selective adaptation showed no diminished effect of adaptation for the speech of the novel speaker and both speakers were equally affected. This finding that the generalisation of selective adaptation was not affected by the visual speech during exposure is in line with results from previous studies that have argued that selective adaptation is a purely auditory phenomenon (Blumstein, Stevens, & Nigro, 1977; Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). Generalisation across speakers thus occurred for both

effects, despite the availability of speaker information in the audiovisual speech during exposure. It is therefore acoustic similarity, not the absence of speaker identity information in the input that allows for generalisation to occur across speakers. Changes in the perceptual system thus occur after exposure to ambiguous and unambiguous speech, and they show that the system can flexibly adjust to the input it is given.

The experiments in Chapter 2 and 3 addressed changes that occur in the perceptual system at the prelexical level. In Chapter 4, we investigated the nature of the lexical representations that are stored in the mental lexicon and used for the recognition of words (Goldinger, 1996, 1998; McClelland & Elman, 1986; Norris & McQueen, 2008). We were interested in whether speaker-specific information obtained from speech was stored in the mental lexicon together with the lexical representations. We further conducted the experiments in Chapter 4 to establish whether the lexical representations are separate for both speech modalities (as shown to be the case for the phonetic categories in Chapter 2) or shared between the modalities (i.e., amodal). In Chapter 4, long-term cross-modal priming was used to provide new evidence about the modality specificity of lexical representations. Word repetition priming effects across modalities would indicate that the lexical representations are shared, since earlier processing in one modality affects the later processing in another modality. Additionally, these underlying representations could either be episodically detailed and contain specific information about the surface details of previous utterances (Church & Schacter, 1994; Goldinger, 1996, 1998; Schacter & Church, 1992) or they could be abstract and contain only information about the canonical word forms (Johnson, 2005; Ladefoged & Broadbent, 1957; Luce & Lyons, 1998). However, a mixture between abstract and episodic information could also be possible, similar to what has previously been proposed for auditory word recognition (McLennan & Luce, 2005; McLennan, Luce, & Charles-Luce, 2003). The purely abstract and purely episodic accounts predict different outcomes of speaker repetition effects on cross-modal priming. If details about

previously perceived utterances that are stored in long-term memory are amodal, episodic theories would predict that words repeated by the same speaker show a larger effect of repetition priming than words repeated by a different speaker because same-speaker repetitions match up better with the previously perceived episode. Abstract models, on the other hand, predict similarly sized effects of repetition priming for both same-speaker repetitions and different-speaker repetitions assuming that the normalisation process is not word specific.

In order to investigate the specificity of the underlying representations, in Experiment 4.1, listeners first identified auditory-only words during the exposure phase and subsequently identified visual-only words during the test phase. The words that listeners identified in the test phase were either repeated from the exposure phase or new. The speaker that listeners saw producing the visual-only words was either the same speaker that listeners had heard during exposure or a novel speaker. The results of Experiment 4.1 revealed a word repetition priming effect in this long-term priming paradigm despite a change in the modality between presentations (i.e., cross-modal priming). Listeners showed improved recognition of visual-only repeated words presented over new words even though the initial presentation of the word was auditory. Exposure to auditory speech can thus influence the subsequent processing of visual speech. The fact that auditory-to-visual repetition priming was observed using an identification task demonstrated that the effect had a phonological locus. The effect of cross-modal word repetition priming was not affected by speaker repetition and there was no increase in the priming effect for words repeated by the same speaker. This means that the lexical representations were abstract and did not contain speaker-specific information. In Experiment 4.2, we tested whether word repetition and speaker repetition affected explicit memory, which involved participants making an explicit judgement about whether a word had been heard before. Such an explicit memory task may thus be more susceptible to repetitions produced by the same speaker. In Experiment 4.2, listeners performed the same identification tasks in the exposure phase, but in the

test phase they first performed an explicit recognition memory task on each trial before providing their visual-only identification responses. The results of Experiment 4.2 replicated the results of cross-modal repetition priming on visual-only speech identification and extended these findings by showing that neither word repetition nor speaker repetition affected listeners' explicit memory of repeated words. Neither listeners' implicit memory nor their explicit memory was affected by whether or not repetitions came from the same speaker. Together these results mean that auditory speech and visual speech share the same lexical representations. These *amodal* lexical representations are abstract and do not contain information about the surface details of previous utterances. The absence of a speaker repetition effect in this cross-modal priming paradigm suggests that it is unlikely that the incoming speech is compared against stored amodal episodic details about previously perceived utterances. These findings cannot, however, make claims about an interpretation of episodic theories in which the details that are stored are modality-specific.

2. Conclusions

The results of the experiments discussed in Chapters 2 through 4 provide important new insights into the changes that occur in the perceptual system of the listener after exposure to auditory and visual speech. They also provide new evidence on how these changes affect the subsequent processing of speech. The results demonstrate how flexible the perceptual system of the speaker truly is. Listeners continually adjust their perceptual system in order to facilitate the processing of speech and do so automatically. These changes that are made in the perceptual system occur regardless of whether the speech input is problematic. Even when presented with speech that is relatively easy to understand, there are still processes in play that are specifically designed to facilitate recognition.

2.1. Generalisation of speaker-specific information

Certain changes in the perceptual system of the speaker are more broadly applicable than others. The retuning of phonetic categories for a specific modality that occurs on the basis of one speaker's input can affect the subsequent processing of speech produced by a different speaker. Previous work on generalisation of auditory speaker-specific knowledge suggested that transfer is determined by acoustic similarity and/or whether or not speaker identity information was available in the input (Kraljic & Samuel, 2006). In Chapter 3, we demonstrated that speaker-specific knowledge generalised across speakers for acoustically similar sounds, even if information about the identity of the speaker was provided during exposure. This generalisation across speakers depends on the acoustic similarity of the idiosyncrasy for the two speakers and not on the availability of speaker identity information in the input, a distinction that was confounded in the study by Kraljic and Samuel (2006). In other words, listeners are able to apply specific changes in their perceptual system in the processing of speech produced by another speaker as long as that speaker produces acoustically similar idiosyncratic sounds. Listeners may be clearly aware that there was a change in the identity of the speaker but that did not matter for the generalisation of speaker-specific information. The fact that every speaker's idiolect is unique does not necessarily mean that every idiosyncrasy a speaker produces in his or her idiolect is also unique. It is hence advantageous that listeners can reuse adjustments made to a speaker of a particular dialect whenever they perceive similar sounding speech from another speaker with the same dialectal background. Seeing the exposure speaker does not prevent this transfer. The generalisation of phonetic retuning thus clearly reflects a process that is meant to streamline the recognition of speech. These findings are further in line with previous findings that have shown generalisation of phonetic retuning across words and across syllable positions (Jesse & McQueen, 2011; McQueen, Cutler, & Norris, 2006).

In Chapter 3, we demonstrated that speaker knowledge generalised across speakers. In Chapter 2, on the other hand, we found that speaker knowledge did not

generalise across modalities at the prelexical level. Changes in the boundaries between two phonetic categories for one modality do not automatically result in changes to the boundaries between the phonetic categories of another modality. The fact that speaker information is not generalised across modalities indicates that the phonetic categories are not linked and do not get retuned in tandem. Changes in the auditory categories can therefore not influence listeners' expectations for visual-only speech. The failure to generalise further means that listeners will only retune their phonetic categories if they are presented with an idiosyncrasy alongside a disambiguating source of information. Previous studies had always presented the auditory or visual idiosyncrasy together with the disambiguating information, whether it was presented visually or whether it was due to lexical knowledge (Baart & Vroomen, 2010; Bertelson et al., 2003; Norris et al., 2003). The results of Chapter 3 show that this seems to be a necessary condition for phonetic retuning to occur; there must be some existing knowledge that enables rapid interpretation of the ambiguity. This interpretation is similar to that of Jesse and McQueen (2011), who found no phonetic retuning after exposure to ambiguities in word-initial position and argued that this was due to the lack of disambiguating lexical information at the beginning of the word. In our study, listeners had no reason to expect a shift in their phonetic category boundaries for a particular modality when they are not presented with an idiosyncrasy in that modality in the speech input. In other words, despite the fact that listeners heard that the auditory speech contained an idiosyncrasy, they were not shown that this auditory idiosyncrasy had a visual parallel and thus there was no explicit indication that the visual phonetic categories had to be retuned. Together with the results from Chapter 2, these findings suggest that speaker information is stored at the prelexical level but that the information that is stored is specific to the modality from which it was obtained.

2.2. Unidirectional generalisation across modalities

Previous research has demonstrated that the perception of visual-only speech can facilitate the subsequent recognition of auditory-only speech from the same speaker (Rosenblum, 2008; Rosenblum, Miller, & Sanchez, 2007). The results of Chapter 2 and Chapter 4 failed to find effects of generalisation of speaker information across modalities. Taken together, these findings may suggest that the generalisation of speaker information across modalities is unidirectional. The movements of the articulators that make up the visual speech signal provide information about the acoustics of the sound. Given the link between visual speech movements and the resulting audible speech signal (Yehia, Rubin, & Vatikiotis-Bateson, 1998), listeners may be able to use the visual-only speech input to adjust their expectations about a speaker's auditory speech. Familiarity to a speaker through visual-only speech may thus provide sufficient information for listener to learn about a speaker's auditory idiosyncrasies. But the reverse does not seem to be the true. Auditory information does not seem to contain sufficient gestural information for listeners to retune their visual categories to a speaker. Exposure to auditory speech is not very informative about how speakers produce sounds visually. This would contradict theories that posit that listeners are able to extract information about the vocal tract of the speaker from auditory speech (Fowler, 1986, 1991; Fowler, Brown, Sabadini, & Welhing, 2003; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985). If listeners were able to retrieve gestural information from auditory speech, then cross-modal transfer of speaker-specific information should have also been found from auditory to visual speech.

2.3. Shared underlying representations

Listeners appear to make use of shared underlying representations at the lexical level of processing (Buchwald, Winters, & Pisoni, 2009; Dodd, Oerlemans, & Robinson, 1989; Kim, Davis, & Krins, 2004). Words that are repeated are processed

significantly faster and more accurately than words that have not been previously perceived and it does not matter whether the repetitions occur in the same modality of speech. The results in Chapter 4 showed that the processing of auditory-only speech and visual-only speech involve the same representations in the mental lexicon of the listener. Observing effects of word repetition priming on visual-only identification following auditory-only exposure in an identification task further established that the effect has a phonological locus rather than a semantic locus, an issue that remained unclear from the only previous study that looked at auditory-to-visual priming (Dodd et al., 1989). The shared lexical representations are not adjusted to specific speakers and do not contain speaker-specific information. That is not to say, however, that speaker-specific information and other details about the utterance are not retained at all. Recall that the findings of Chapters 2 and 3 demonstrated that at the prelexical level speaker information was indeed retained, albeit in a modality-specific way. It does, however, indicate that speaker information is not involved once contact is made with the lexical representations found in long-term memory. The shared lexical representations thus constitute abstract, canonical forms of words stored in the mental lexicon. At the lexical level, the information that is used by listeners is thus amodal, and speaker information is not encoded at this level.

2.4. A way forward...

The present thesis has provided a number of important new findings that add to our ever-growing understanding of how listeners interpret language and how they adjust their perceptual system to the speech with which they are presented. Listeners are able to make adjustments to auditory speech that is ambiguous and therefore problematic. But even the processing of unambiguous speech can result in changes in the perceptual system when listeners repeatedly hear the same sound being presented. Both auditory speech and visual speech are prevalent in everyday communication and listeners are able to adjust to idiosyncrasies that occur in both modalities separately either using additional knowledge about the language or

information that is simultaneously presented in another modality. At the prelexical level of processing, listeners are able to use information obtained about the speech from a particular speaker in order to facilitate the subsequent recognition of speech produced by that same speaker. This is only the case, however, when the speech is presented in the same modality on both occasions but does not affect the interpretation of speech across modalities. In certain cases, speaker information that is stored after exposure to one speaker can aid even the prelexical processing of speech from another speaker. In such cases, acoustic similarity in the speech input from the two speakers means that listeners can reapply previously retuned phonetic categories to the processing of speech from a novel speaker. Speaker-specific idiosyncrasies to which listeners attune must, however, be presented together with a disambiguating source of information. For instance, listeners cannot adjust visual phonetic categories using information obtained from auditory-only speech. Phonetic categories at the prelexical level are thus separate for auditory and visual speech. It appears that adjustments to speakers occur mainly at the prelexical level, and once the mental lexicon becomes involved, details of the specific utterance that was perceived are no longer of importance. The lexical representations that are stored in listeners' mental lexicon are shared between the modalities, however, and the same representations are involved in the processing of auditory and visual speech. Speaker-specific information is not stored amodally together with lexical representations, and words show the same improvement of having been previously presented regardless of who produced them. Together, the results of the experimental chapters show that speaker information is stored prelexically but is not encoded in the mental lexicon and that lexical representations in long-term memory are amodal but the phonetic categories that are used at the prelexical level are modality-specific. Generalisation of speaker knowledge is thus possible across words but not across modalities. These findings therefore show that our perceptual system is not rigid but rather is highly adaptable and can overcome many interpretation problems simply by fine-tuning certain predefined settings.

Bibliography

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of the Acoustical Society of America, 126*(5), 2649-2659.

Ades, A. E. (1974). How phonetic is selective adaptation: Experiments on syllable position and vowel environment. *Perception & Psychophysics, 16*(1), 61-66.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113*(1), 544-552.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339-355.

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

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The Celex lexical database* [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes [Software]. Available from http://lme4.r-forge.r-project.org/.

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1-4), 5-18.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62*(2), 233-252.

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

Blumstein, S. E., Stevens, K. N., & Nigro, G. N. (1977). Property detectors for bursts and transitions in speech perception. *Journal of the Acoustical Society of America, 61*(5), 1301-1313.

Boersma, P., & Weenink, D. (2006). Praat: Doing phonetics by computer (Version 5.2.40) [Software]. Available from http://www.praat.org/

Bond, Z. & Moore, T. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, *14*, 325-337.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America, 84*(1), 101-114.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206-219.

Breeuwer, M., & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America, 79*(2), 481-499.

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes, 24*(4), 580-610.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 521-533.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*(6), 3647-3658.

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26*, 274-284.

Creelman, C. D. (1957). The case of the unknown talker. *Journal of the Acoustical Society of America, 29*, 655.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). *Prelexically-driven perceptual retuning of phoneme boundaries.* Paper presented at the Proceedings of Interspeech: 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia.

Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition, 122*(3), 442-453.

Daly, N., Bench, R. J., & Chappell, H. (1997). Gender differences in visual speech variables. *Journal of the Academy of Rehabilitative Audiology, 30*, 63-76.

Davis, C., & Kim, J. (2006). Audio-visual speech off the top of the head. *Cognition, 100*(3), B21-B31.

Diehl, R. L. (1975). Effect of selective adaptation on identification of speech sounds. *Perception & Psychophysics, 17*(1), 48-52.

Dodd, B., Oerlemans, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lipread, graphic and hear stimuli. *Visible Language, 22*(1), 58-77.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*(1), 99-109.

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

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

Ellis, A. (1982). Modality-specific repetition priming of auditory word recognition. *Current Psychological Research, 2*, 123-128.

Fant, G. (1973). *Speech sounds and features*. Cambridge, MA: MIT Press.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*, 796-804.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics, 34*(4), 409-438.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics, 14*(1), 3-28.

Fowler, C. A. (1991). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99*(3), 1730-1741.

Fowler, C. A., Brown, J. M., Sabadini, L., & Welhing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49*(3), 396-413.

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

Gagné, J.-P., Masterson, V.M., Munhall, K.G., & Bilida, N. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology 27, 135-158.*

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Grant, K. W., & Seitz, P. F. (2000a). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *Journal of the Acoustical Society of America, 107*(2), 1000-1011.

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

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*(5), 2677-2690.

Hadar, U., Steiner, T.J., Grant, E.C., & Clifford Rose, F. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science, 2*, 35-46.

Hadar, U., Steiner, T.J., Grant, E.C., & Clifford Rose, F. (1984). The timing of shifts of head postures during conversation. *Human Movement Science, 3*, 237-245.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America, 117*(2), 842-849.

Irwin, J. R., Whalen, D. H., & Fowler, C. A. (2006). A sex difference in visual influence on heard speech. *Perception & Psychophysics, 68*(4), 582-592.

Jackson, A., & Morton, J. (1984). Facilitation of auditory word recognition. *Memory & Cognition, 12*(6), 568-574.

Jesse, A. (2005). *Towards a fuzzy logical model of perception: The time-course of information in lexical identification of face-to-face speech.* University of California, Santa Cruz.

Jesse, A., & Janse, E. (2009). Seeing a speaker's face helps stream segregation for younger and elderly adults. *Journal of the Acoustical Society of America, 125*(4), 2361.

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics, 72*(1), 209-225.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review, 18*(5), 943-950.

Jesse, A., McQueen, J. M., & Page, M. (2007). The locus of talker-specific effects in spoken word recognition. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1921-1924). Dudweiler: Pirrot.

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000/2001). The processing of information from multiple sources in simultaneous interpreting. *Interpreting, 5*(2), 95-115.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363-389). Malden, MA: Blackwell.

Jones, B. C., Feinberg, D. R., Bestelmeyer, P. E. G., DeBruine, L. M., & Little, A. C. (2010). Adaptation to different mouth shapes influences visual perception of ambiguous lip speech. *Psychonomic Bulletin & Review, 17*(4), 522-528.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46*(2), 390-404.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': Matching identity across modality. *Current Biology, 13*(19), 1709-1714.

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication, 27*(3), 187-207.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition, 93*(1), B39-B47.

Krahmer, E., Ruttkay, Z., Swerts, M., & Wesselink, W. (2002). Pitch, eyebrows, and the perception of focus. *SP-2002*, 443-446.

Krahmer, E. J., & Swerts, M. (2004). More about brows: A cross-linguistic study via analysis-by-synthesis. In Z. Ruttkay & C. Pelachaud (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents* (pp. 191-216). Dordrecht: Kluwer Academic Publishers.

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

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

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

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

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iversion, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13-F21.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56*(3), 485-502.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*(1), 98-104.

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: Matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4), 905-914.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 1-32). Cambridge: Cambridge University Press.

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

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

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 26*(4), 708-715.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1-36.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics, 24*(3), 253-257.

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131-142.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 676-684.

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

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*(5), 753-771.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97*(2), 225-252.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543-562.

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

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

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 306-321.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(4), 539-553.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*(6), 1113-1126.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech, 49*, 101-112.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82*(6), 2145-2147.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology: General, 41*(5), 329-335.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41*(4), 215-225.

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical tone: Evidence from a learning paradigm. *Cognitive Science, 35*(1), 184-197.

Mitterer, H., & De Ruiter, J. P. (2008). Recalibrating color categories using word knowledge. *Psychological Science, 19*(7), 629-634.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*(4), 379-390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*(1), 365-378.

Munhall, K.G., & Buchan, J.N. (2004). Something in the way she moves. *Trends in Cognitive Science, 8*(2), 51-53.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15*(2), 133-137.

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science, 13*(3), 407-420.

Norris, D., Butterfield, S., McQueen, J. M., & Cutler, A. (2006). Lexically guided retuning of letter perception. *Quarterly Journal of Experimental Psychology, 59*(9), 1505-1515.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357-395.

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

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355-376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*(1), 42-46.

Orfanidou, E., Davis, M. H., Ford, M. A., & Marslen-Wilson, W. D. (2011). Perceptual and response components in repetition priming of spoken words and pseudowords. *Quarterly Journal of Experimental Psychology, 64*(1), 96-121.

Orfanidou, E., Marslen-Wilson, W. D., & Davis, M. H. (2006). Neural response suppression predicts repetition priming of spoken words and pseudowords. *Journal of Cognitive Neuroscience, 18*(8), 1237-1252.

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28*, 381-393.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 309-328.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175-184.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication, 13*(1-2), 109-125.

R Development Core Team (2007). R: A language and environment for statistical computing [Software]. Available from http://www.R-project.org/.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 97-113). London, U.K.: Lawrence Erlbaum.

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

Rosenblum, L. D. (2005). The primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78). Malden, MA: Blackwell.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405-409.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science, 18*(5), 392-396.

Rosenblum, L. D., Yakel, D. A., & Green, K. P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 806-819.

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

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

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

Sawusch, J. R., & Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. *Journal of Experimental Psychology: Human Perception and Performance, 7*(2), 408-421.

Sawusch, J. R., & Pisoni, D. B. (1976). Response organization in selective adaptation to speech sounds. *Perception & Psychophysics, 20*(6), 413-418.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 915-930.

Scharenborg, O., Mitterer, H., & McQueen, J. M. (2011). *Perceptual learning of liquids.* Paper presented at the Proceedings of Interspeech: 12th Annual Conference of the International Speech Communication Association, Florence, Italy.

Sheffert, S. M. (1998). Voice-specificity effects on auditory word priming. *Memory & Cognition, 26*(3), 591-598.

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language, 34*(5), 665-685.

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

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition, 92*(3), B13-B23.

Spehar, B. P., Tye-Murray, N., & Sommers, M. S. (2008). Intra- versus intermodal integration in young and older adults. *Journal of the Acoustical Society of America, 123*(5), 2858-2866.

Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2009). Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. *Neuropsychologia, 47*, 972-979.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212-215.

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). London, U.K.: Lawrence Erlbaum.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences, 335*(1273), 71-78.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 2*(3), 339-363.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352-373.

Van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *Journal of the Acoustical Society of America, 134*(1), 562-571.

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

Van Son, N. J. D. M. M., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1994). Viseme classifications of Dutch consonants and vowels. *Journal of the Acoustical Society of America, 96*(3), 1341-1355.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926-940.

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

Vroomen, J., & Baart, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech, 52*, 341-350.

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

Vroomen, J., Van Linden, S., Keetels, M., De Gelder, W., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication, 44*(1-4), 55-61.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, J. C. (1977). Effects of Training on the Visual Recognition of Consonants. *Journal of Speech and Hearing Research, 20*, 130-145.

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech and Hearing Research, 17*(2), 270-278.

Webster, M. A. (2004). Pattern-selective adaptation in color and form perception. In L. Chalupa & J. Werner (Eds.), *The Visual Neurosciences* (pp. 936-947). Cambridge, MA: MIT Press.

Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review, 6*(4), 647-653.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49-63.

Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics, 62*(7), 1405-1412.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1-2), 23-43.

Nederlandse samenvatting

Iedere dag word je als luisteraar blootgesteld aan een groot aantal verschillende sprekers met ieder zijn of haar eigen taalgebruik en spraakpatroon. Zo kan iemand bijvoorbeeld een accent hebben, waardoor deze persoon bepaalde klanken net even anders uitspreekt dan je gewend bent. Na een tijdje naar dezelfde spreker geluisterd te hebben, lijkt alles vanzelf een stuk soepeler te gaan. Op basis van ervaring met de klanken van een bepaalde spreker vinden er namelijk processen plaats in je brein die ervoor zorgen dat je deze persoon op den duur makkelijker kunt verstaan. In mijn onderzoek kijk ik naar een aantal van deze aanpassingen en hoe ze invloed kunnen hebben op de manier waarop een volgend spraaksignaal wordt geïnterpreteerd. Ik heb hierbij gebruik gemaakt van auditieve spraak (het stemgeluid van de spreker), maar ook van visuele spraak (de mondbewegingen van de spreker) en audiovisuele spraak (de combinatie van beide bronnen). Hoewel het mogelijk is om iemand puur op basis van het stemgeluid te verstaan, vormt het visuele spraaksignaal een belangrijke bron van informatie. Dit is vooral erg duidelijk wanneer iemand lastig is te verstaan. Wanneer er veel achtergrondgeluid is (bijvoorbeeld in een café), kan het zien van de mondbewegingen je helpen de spreker te verstaan. Dit wil echter niet zeggen dat de visuele informatie alleen dient als een soort van back-up. Wanneer aanwezig, wordt het visuele spraaksignaal automatisch door de luisteraar geïnterpreteerd en wat je de spreker hoort zeggen wordt beïnvloed door het visuele signaal. In dit proefschrift heb ik nieuwe kennis vergaart over hoe luisteraars zich aanpassen aan een spreker op basis van wat ze horen en wat ze zien. Onderzoek naar (audio)visuele spraakherkenning is belangrijk om een volledig beeld te krijgen van wat er voor nodig is om een spreker te kunnen verstaan. Een beschrijving van wat er voor nodig is om een spreker te kunnen verstaan zou daarom dan ook niet compleet zijn als het visuele aspect zou worden genegeerd.

In Hoofdstuk 2 en 3 kijk ik specifiek naar een proces dat *phonetic retuning* wordt genoemd. Dit is een proces vindt plaats op het niveau van de individuele klanken. In je brein heb je zogenaamde fonetische categorieën die je gebruikt om de klanken in het inkomende spraaksignaal te analyseren en beoordelen. Deze fonetische categorieën bestaan zowel voor het auditieve spraaksignaal als voor het visuele spraaksignaal. Je beoordeelt dus welke klank een spreker heeft geproduceerd door te kijken in welk hokje het waargenomen geluid of de mondbeweging het beste past. De grenzen die je fonetische categorieën afbakenen zijn grotendeels bepaald door je moedertaal, maar ze blijven flexibel zodat ze aangepast kunnen worden wanneer nodig. Met *phonetic retuning* wordt het bijstellen of herkalibreren van deze grenzen bedoeld. Zo kan het zijn dat je een spreker tegenkomt die een klank produceert die eigenlijk precies tussen twee categorie in lijkt te liggen. In plaats van een duidelijke f-klank of een duidelijke s-klank zegt deze spreker bijvoorbeeld altijd een klank die ergens tussen de twee in lijkt te liggen. Dit kan erg verwarrend zijn en om dit probleem op te lossen wordt de grens tussen de twee categorieën in kwestie wat bijgeschoven, waardoor je uiteindelijk de rare klank alsnog in de juiste categorie kunt plaatsen. Een dergelijke verschuiving van je categoriegrenzen gebeurt alleen wanneer je extra informatie beschikbaar hebt waaruit blijkt welke klank de spreker eigenlijk had bedoeld. Een van de manieren waarop dit kan gebeuren is met behulp van je woordkennis. Zo zal de vreemde klank tussen s en f (aangeduid met ?) makkelijk te interpreteren zijn wanneer deze alleen voorkomt in woorden als *olij?* en *gira?*, doordat in deze context alleen de f-klank mogelijk is. Als je dezelfde klank herhaaldelijk tegenkomt in een dergelijke context zul je leren dat deze spreker de f gewoonweg wat raar uitspreekt. Op basis van deze informatie wordt de grens tussen de f-categorie en de s-categorie dan wat opgeschoven, zodat de rare klank van deze spreker daarna kan worden ingedeeld binnen de juiste categorie. De grens zal de andere kant op worden geschoven als je de rare klank alleen maar tegen komt in woorden met een s-context (bijvoorbeeld *radij?*). Dit proces van *phonetic retuning*

heeft geen effect op hoe een luisteraar een klank hoort, alleen op hoe deze klank wordt beoordeeld.

In Hoofdstuk 2 laat ik zien dat vergelijkbare verschuivingen in de categoriegrenzen ook plaatsvinden voor je visuele fonetische categorieën. Ook de mondbewegingen die iemand produceert kunnen namelijk ambigu zijn tussen twee mogelijke klanken. Uit eerder onderzoek is gebleken dat wanneer je een ambigue klank ziet en tegelijkertijd een normale, canonieke klank hoort, je de auditieve informatie kunt gebruiken om je visuele categoriegrens bij te stellen. De resultaten in Hoofdstuk 2 tonen voor het eerst aan dat ook woordkennis gebruikt kan worden voor het aanpassen van de visuele categorieën. Dit resultaat geeft dus aan dat ook de visuele categorieën kunnen worden bijgesteld aan de hand van informatie die al in je hoofd is opgeslagen en dat extra informatie vanuit je andere zintuigen hiervoor niet per se nodig is. Je kunt je categorieën dus aanpassen om het probleem op te lossen zolang je kunt interpreteren wat de bedoelde woordcontext was waarin de rare visuele klank voorkwam. Ook onderzoek ik in Hoofdstuk 2 of de fonetische categorieën die gebruikt worden bij het interpreteren van auditieve en visuele spraak met elkaar in verbinding staan. Het zou zo kunnen zijn dat verschuivingen in de grens tussen twee auditieve categorieën ook worden toegepast op de plaatsing van de grens tussen de gerelateerde visuele categorieën. Doordat de mondbeweging van de spreker grotendeels beïnvloedt hoe het uiteindelijke geluid zal klinken, kan het zo zijn ambigue klank voortkomt uit een ambigue mondbeweging. Het zou in dat geval handig zijn als de oplossing van het auditieve probleem in de ene instantie ook meteen zou zorgen voor dezelfde oplossing voor het visuele probleem. Dit zou tevens aangeven dat de luisteraar zich er niet direct van bewust hoeft te zijn dat er een probleem optreedt in een van de twee signalen. De resultaten in Hoofdstuk 2 wijzen er echter op dat de fonetische categorieën voor audio en video niet met elkaar in verbinding staan. Een verandering in de grens tussen twee auditieve categorieën heeft namelijk geen invloed op de interpretatie van het visuele spraaksignaal. Hierbij blijkt dan ook dat de grens tussen de visuele categorieën niet is verschoven. Een

probleem dat geconstateerd werd op basis van wat de luisteraar hoorde werd opgelost, maar toen eenzelfde probleem zich daarna voordeed in het visuele spraaksignaal was dit niet automatisch al verholpen. De grenzen tussen fonetische categorieën moeten dus apart worden aangepast voor auditieve spraak en visuele spraak. Het proces van *phonetic retuning* lijkt daardoor dan ook alleen plaats te vinden wanneer het voor de luisteraar expliciet duidelijk is dat er zich een probleem voordoet.

In Hoofdstuk 3 kijk ik naar auditieve fonetische categorieën en onderzoek ik of veranderingen in de categoriegrenzen voor een spreker ook effect kunnen hebben op de interpretatie van het spraaksignaal van een andere spreker die een vergelijkbare klank produceert. Met andere woorden, worden aanpassingen in de grenzen tussen categorieën specifiek voor een bepaalde spreker gedaan of zijn deze aanpassingen globaal toepasbaar? Als een aanpassing kan generaliseren naar het spraaksignaal van andere sprekers zou dat natuurlijk handig zijn bij het verstaan van een nieuwe spreker met hetzelfde accent. De hoofdvraag in Hoofdstuk 3 is echter nog wat specifieker. Ik test namelijk of de veranderingen in categoriegrenzen kunnen worden toegepast bij het verstaan van een andere spreker wanneer het voor de luisteraar vrij duidelijk is dat het inderdaad een andere spreker betreft. De resultaten van Experiment 3.1 geven aan dat een verschuiving in de grens tussen twee auditieve categorieën ook effect heeft op de interpretatie van het spraaksignaal van een andere spreker, mits de tweede spreker een vergelijkbare klank produceert. Dit laat zien dat dezelfde oplossing kan worden gebruikt voor hetzelfde probleem bij twee verschillende sprekers. Het proces van het verschuiven van de grenzen hoeft niet opnieuw te worden doorlopen voor de tweede spreker. Bovendien maakt het niet uit voor de luisteraar dat de identiteit van de twee sprekers anders is. Het enige dat telt is dat ze akoestisch vergelijkbaar zijn. In Hoofdstuk 3 laat ik verder nog zien dat een geheel ander proces, waarbij luisteraars minder gevoelig worden voor een geluid dat herhaaldelijk wordt gepresenteerd (*selective adaptation*), ook de interpretatie van het spraaksignaal van zowel de oorspronkelijke spreker als een

andere spreker kan beïnvloeden. Deze resultaten samen met die van Hoofdstuk 2 laten zien dat het brein van de luisteraar zeer flexibel is en dat bepaalde instellingen snel kunnen worden aangepast om potentiële problemen in het inkomende spraaksignaal te op te lossen. Aanpassingen kunnen worden gedaan met welke informatie er op dat moment voorhanden is en kunnen vervolgens worden toegepast voor andere, vergelijkbare sprekers. Bovendien wordt het steeds duidelijker dat bepaalde effecten die eerder alleen waren gevonden voor auditieve spraak ook plaatsvinden voor visuele en audiovisuele spraak.

In Hoofdstuk 4 zoom ik uit van het klankniveau naar het woordniveau en bekijk ik hoe woordkennis is opgeslagen in het brein van de luisteraar. Je woordkennis bevindt zich in je mentale lexicon en deze kennis wordt gebruikt om woorden te vormen uit de waargenomen individuele klanken. Zo bekijk ik in Hoofdstuk 4 of er specifieke informatie over sprekers wordt opgeslagen in je mentale lexicon. Een van de mogelijkheden is dat de gegevens die je bewaart veel informatie bevatten over de verschillende vormen van een woord die je eerder hebt waargenomen en dat je bij het interpreteren van spraak het binnenkomende spraaksignaal vergelijkt met deze eerdere vormen. In dat geval bewaar je dus veel specifieke informatie over sprekers samen met de woordvormen. Een andere mogelijkheid is dat je juist alle overbodige informatie van een specifieke eerder gehoorde uitspraak weggooit en dat alleen de abstract, canonieke woordvorm ligt opgeslagen in je lexicon. Ook bekijk ik in Hoofdstuk 4 of de onderliggende woordvormen in je mentale lexicon gescheiden zijn voor auditieve en visuele spraak (zoals bleek voor de fonetische categorieën), of dat er voor beide bronnen slechts één woordvorm is opgeslagen die informatie bevat voor zowel horen als liplezen. Wat ik zie in Hoofdstuk 4 is dat je woorden die je kort daarvoor hebt gehoord vervolgens beter kunt liplezen dan woorden die je nog niet hebt gehoord. Een dergelijk effect van herhaling van woorden, waarbij het woord in tweede instantie makkelijker te interpreteren is, wordt ook wel *repetition priming* genoemd. Het *priming*-effect ontstaat doordat een woordvorm bij de eerste presentatie wordt aangeroepen en

daardoor vervolgens bij de tweede presentatie makkelijker te vinden is. Het *priming*-effect is al een lange tijd bekend, maar is tot dusverre voornamelijk getest in situaties waarbij zowel de eerste als tweede herhaling van het woord auditief werd gepresenteerd. Mijn resultaten laten echter zien dat een eerste auditieve presentatie vervolgens een positief effect kan hebben op het liplezen van hetzelfde woord. Dit wijst erop dat zowel auditieve als visuele spraaksignalen gebruik maken van dezelfde woordvormen in het mentale lexicon. Als er gescheiden woordvormen zouden zijn opgeslagen voor auditieve en visuele spraak zou een eerste auditieve presentatie geen herhalingseffect kunnen hebben op de tweede, visuele presentatie. Ook geven mijn resultaten aan dat er geen specifieke informatie over de spreker wordt opgeslagen met de woordvormen in het mentale lexicon. Was dit het geval geweest, dan zou een herhaling door dezelfde spreker makkelijker te liplezen zijn geweest dan een herhaling door een andere spreker, doordat een herhaling door dezelfde spreker sterker zou lijken op de eerder aangeroepen woordvorm dan een herhaling door een andere spreker. Dit is niet wat ik zie in mijn resultaten: deelnemers kunnen herhaalde woorden beter liplezen dan nieuwe woorden, ongeacht of de spreker hetzelfde is of verandert in de twee keren dat het woord wordt gepresenteerd. Het *priming*-effect wordt dus niet (positief) beïnvloed door de identiteit van de spreker en daaruit blijkt dat bij het interpreteren van een eerdere presentatie geen specifieke informatie is opgeslagen over hoe de spreker klonk. Het lijkt er dus op dat auditieve en visuele spraaksignalen wel gebruik maken van dezelfde woordvormen, maar dat deze woordvormen gereduceerd zijn tot abstracte, canonieke weergaven.

De drie experimentele hoofdstukken in mijn proefschrift geven nieuwe inzichten in de informatie die is opgeslagen in het brein van de luisteraar en ook in hoe deze informatie kan worden bijgewerkt na het horen en zien van spraak. Tevens geven de resultaten nieuwe informatie over hoe veranderingen in het brein invloed kunnen hebben op de interpretatie van spraak die door een totaal andere spreker wordt geproduceerd. Informatie over sprekers wordt bewaard op klankniveau, waar

de informatie die is opgeslagen apart is voor auditieve en visuele spraak, maar informatie over de spreker is minder belangrijk op het woordniveau, waar we juist zien dat dezelfde informatie wordt gebruikt voor zowel horen als liplezen. Dit alles geeft aan dat het gehele systeem dat een rol speelt bij het begrijpen van spraak niet vast staat, maar juist enorm flexibel is en kan worden bijgewerkt om problemen op te lossen.

Curriculum Vitae

Patrick van der Zande was born in 1986 in 's-Hertogenbosch, the Netherlands. He studied English Language & Culture, with a specialisation in linguistics, at the Radboud University Nijmegen. In 2006, he spent a six-month period at Lund University in Sweden under the Erasmus programme. He received a bachelor's degree in 2007, followed by a master's degree (cum laude) in 2009. In the same year, he also obtained a master's degree in Language & Communication, after completing a one-year research master co-organised by the Radboud University Nijmegen and Tilburg University. He was awarded a three-year scholarship from the Max Planck Society to do a PhD at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands and joined the Language Comprehension Group. During his three-year Ph.D. track, he also spent four months working in the United States at the University of Massachusetts, Amherst. Currently, he is looking for a challenging new career path.

List of publications

Van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *Journal of the Acoustical Society of America, 134*(1), 562-571.

Van der Zande, P., Jesse, A., & Cutler, A. (Under revision). Hearing words helps seeing words: A cross-modal word repetition effect. *Speech Communication.*

Van der Zande, P., Jesse, A., & Cutler, A. (Under revision). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics.*

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout.*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*.
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*.
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*.
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*.
6. Gesture and speech production. *Jan-Peter de Ruiter.*
7. Comparative intonational phonology: English and German. *Esther Grabe*.
8. Finiteness in adult and child German. *Ingeborg Lasser*.
9. Language input for word discovery. *Joost van de Weijer*.
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*.
11. Producing past and plural inflections. *Dirk Janssen*.
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*.
13. From speech to words. *Arie van der Lugt*.
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*.
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*.
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*.
17. Moving eyes and naming objects. *Femke van der Meulen*.

18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*.
19. Morphology in speech comprehension. *Kerstin Mauth*.
20. Morphological families in the mental lexicon. *Nivja H. de Jong*.
21. Fixed expressions and the production of idioms. *Simone A. Sprenger*.
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*.
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*.
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*.
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*.
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*.
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*.
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*.
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*.
30. A grammar of Jalonke argument structure. *Friederike Lüpke*.
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*.
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*.
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*.
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*.
35. Retrieving semantic and syntactic word properties. *Oliver Müller*.

36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*.
37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*.
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*.
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*.
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich.*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*.
42. The acquisition of auditory categories. *Martijn Goudbeek*.
43. Affix reduction in spoken Dutch. *Mark Pluymaekers*.
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*.
45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*.
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*.
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*.
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*.
49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*.
50. Lexical processing of morphologically complex words: An information theoretical perspective. *Victor Kuperman*.
51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*.
52. Prosodic structure in speech production and perception. *Claudia Kuzla.*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke.*

54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*.
55. Processing the fine temporal structure of spoken words. *Eva Reinisch*.
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*.
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*.
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*.
59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*.
60. Processing casual speech in native and non-native language. *Annelie Tuinman*.
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*.
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*.
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*.
64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*.
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*.
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*.
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*.
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*.
69. Structuring language: contributions to the neurocognition of syntax. *Katrien R. Segaert*.
70. Infants' appreciation of others' mental states in prelinguistic communication: a second person approach to mindreading. *Birgit Knudsen*.

71. Gaze behavior in face-to-face interaction. *Federico Rossano*.
72. Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space. *Connie de Vos*.
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*.
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt J. Witteman*.
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*.
76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*.
77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*.
78. Contributions of executive control to individual differences in word production. *Zeshu Shao*.
79. Hearing and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*.