Lexically-guided perceptual learning in speech processing

ISBN-10 90-76203-22-9
ISBN-13 978-90-76203-22-5

Cover design: Linda van den Akker, Inge Doehring
Cover illustration: Frank Eisner
Printed and bound by Ponsen & Looijen bv, Wageningen

Lexically-guided perceptual learning in speech processing

een wetenschappelijke proeve
op het gebied van de Sociale Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus Prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op dinsdag 7 maart 2006
des namiddags om 1.30 uur precies

door

Frank Eisner

geboren op 9 januari 1976
te Saarbrücken, Duitsland

Promotor:	Prof. dr. A. Cutler
Co-promotor:	Dr. J.M. McQueen
Manuscriptcommissie:	Prof. dr. A.J.W.M. Thomassen
	Dr. P. Indefrey
	Prof. dr. P. Bertelson

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

Für meine Eltern

Dankwoord

So now I am almost at the point where there is nothing more left to do for my dissertation other than to say thanks to all the people who have helped me get here. As always I've left everything to the last minute and the whole book needs to be sent off for printing very soon, so I will try to keep it short, and I need to apologise in advance to everyone I forgot here.

First of all I would like to thank my promotor Anne Cutler for having me in her group at the Max Planck Institute and for her support through the last four years. I thank Anne especially for her contagious enthousiasm for the work we do, which has helped me more than once to find my motivation back at times when things were not going so smoothly. I always appreciated Anne's critical reading and speedy feedback on my writing, and I very much enjoyed the exciting and inspiring atmosphere at work, which I think is due largely to Anne's style of running the Comprehension group.

Second, I cannot thank my co-promotor James McQueen enough for his day-to-day input on my project and just general, plain invaluable support. I could always be certain that James was thinking along with me, if not ahead of me, about what I was up to and where my work was going. At the same time he got me to always try harder, in a good way. Most of what I know now about how to work and think scientifically I owe to James' supervision, and I am extremely grateful for that.

I enjoyed the many interesting discussions with the members of the Comprehension group, and I thank Roel Smits especially for his insightful comments on my project.

Then there is the whole PhD community at the MPI – there are too many of you to list here, so I just want to thank my PhD colleagues for the friendly, collegial, and fun atmosphere at work and all the social events ouside work. A special thank you to those that I've had the pleasure to share an office with, and I also want to say here that I'm very proud to have Claudia Kuzla and Heidrun Bien as my paranimfen.

When I first arrived at the MPI I got to meet Ethan Cox, Alissa Melinger, Greg Gulrajani & friends, who made me feel at home quickly, introduced me to cool music, games, and were just great to hang out with. Thanks for all that, folks. Life in Nijmegen wasn't the same anymore after you left.

Ik dank Ad Verbunt, Kees van der Veer en John Nagengast voor technische hulp, en ik dank hen samen met het hele volleybalclubje voor veel lol op dinsdagavonden. Verder dank ik alle mensen die zorgen dat het onderzoek doen op het MPI zo soepel afloopt, en hierbij mijn bijzonder dank aan Agnes de Boer en José Wannet voor hun hulp met het inroosteren van proefpersonen, en voor de gezellige praatjes tijdens het wachten als deze dan soms toch niet verschenen.

Half-way through my project I had the priviledge to work with Miranda van Turennout and the Learning and Plasticity group at the FC Donders Centre, who introduced me to the world of functional neuroimaging and without whom chapter 4 of this thesis would not exist. Ik ben Miranda heel dankbaar voor haar leuke begeleiding, voor haar enthousiasme voor het project, en voor alle dingen die ik van haar over fMRI heb geleerd. Ten tweede gaat mijn dank uit naar Jos Olders, die ontzettend veel tijd heeft gestoken in ons project en enorm veel heeft geholpen met de data-analyses. Ook als deze analyses soms frustrerend waren was het werken met Jos altijd gezellig en ik heb er heel veel aan gehad. Op het Donders Centrum dank ik ook Paul Gaalman, Jens Schwarzbach, en Marieke van der Linden voor hun hulp met allerlei technische dingen, Barbara Wagensveld voor haar hulp met de dataverzameling, en Peter Hagoort voor het mogelijk maken van deze samenwerking.

During a 6-month break from the PhD work, I was very lucky to be a visitor in Sophie Scott's research group at the Institute of Cognitive Neuroscience in London. I would like to thank everyone in the Speech Communication group, as well as Richard Wise and Jane Warren at the Hammersmith Hospital, and Stuart Rosen at the Phonetics Department, for a terrific time, both in London and away at conferences. I am thrilled to be back in this group now, and I greatly appreciated the support I had from everyone here during the first months when I was still writing the final pages of this thesis.

Schliesslich möchte ich meiner Schwester Maike danken für alle fröhlichen Besuche, e-mails, und Telefongespräche in den letzten Jahren, und ganz besonders herzlich meinen Eltern, die mich bei allen Entscheidungen immer nur unterstützt haben, und ohne deren Hilfe ich niemals bis hierhin hätte kommen können.

Femke, ten slotte dank ik jou voor al je geduld, liefde, en steun.

London, 10 January 2006

Contents

Chapter 1

Introduction

Speech is about the most complex acoustic signal we encounter on a regular basis. The signal is rich in information that the listener may exploit for decoding the meaning intended by the speaker. At the same time the signal contains non-linguistic information about the speaker, and frequently carries other sounds from the environment. As yet, the ability of the human brain to extract a linguistic message from this signal is unmatched by the performance of computers. The brain draws on highly specialised systems for this task, some of which are relatively static and have developed over the course of evolution or are established early on in life, while others are dynamic and able to adapt rapidly to changing contexts. It is the dynamic nature of parts of the perceptual system which allows us to understand speech effortlessly despite changes in speakers, accents, or background noises — the kind of factors which usually have catastrophic consequences on the performance of computerised speech recognition systems. This thesis aims to contribute to a better understanding of the processes that underlie such rapid adjustments. The focus will be on learning that occurs when listeners encounter a talker who consistently articulates a particular speech sound in an unusual way. There are several issues involved in this research, for example, what the relationship between speech perception and the identity of a talker is, how such rapid perceptual adjustments relate to other types of learning, how well current models of speech recognition can account for this process, and which neural mechanisms might be implicated. In this first chapter, some of the relevant literature concerning those four topics will be reviewed.

1.1 Variability and talker specificity in speech

Much research in speech perception has been devoted to the phenomenon of perceptual constancy: the ability of listeners to perceive speech sounds reliably despite considerable variability in the acoustic signal. The factors underlying this variability are numerous and include speech rate fluctuations, individual differences between talkers' vocal tract shapes, ambient noise, affect, or dialects. To date, no complete set of invariant physical attributes of the speech signal has been found from which the perception of the speech sounds of a language could be reliably predicted. The problem is that two utterances of the same speech sound are extremely unlikely to ever be physically identical, not even when produced by the same talker, and certainly not when produced by different talkers. Worse, physically identical sounds can elicit different phonemic percepts depending on context (Repp & Liberman, 1987). In models of spoken word recognition it is commonly assumed that the perceptual system deals with such variability by extracting relevant information from the signal in a complex normalisation process, details of which are not well understood. The products of the normalisation process are relatively simple abstract units of representation (e.g., phonemes or features) that can be further processed and mapped onto equally abstract symbolic representations of words in the lexicon (Halle, 1985). According to an extreme version of this view, information about voice, dialect, affect, etc. is therefore redundant, discarded in the computations leading to lexical access, and processed by a separate faculty.

Support for the view that perception of words and voices are independent processes is provided by findings suggesting that one function can be isolated from the other. For example, in whispered speech or noise-vocoded speech, information about the identity of the talker is largely lost while comprehension remains fairly effortless (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Accordingly, different acoustic properties of the speech signal are said to carry information about one or the other perceptual function. Further evidence for a functional independence of voice processing and lexical access has come from demonstrations of double dissociations in neuropsychological investigations. In receptive types of aphasia, typically after damage to the left temporal lobe, speech comprehension is often impaired while voice recognition remains intact. A right temporal lobe infarction, in contrast, can produce the reverse pattern of impaired talker recognition ability in the absence of a comprehension deficit (Van Lancker, Cummings, Kreiman, & Dobkin, 1988; Van Lancker, Kreiman, & Cummings, 1989; Peretz et al., 1994).

The nature of the speech signal, however, is such that it carries multiple acoustic cues to a particular speech sound at any given time in parallel, and the perceptual system can tolerate the absence of one or more such cues without too much trouble. Because of this built-in redundancy, the fact that speech can still be understood when it is whispered or artificially manipulated to remove talker identity information does not mean that the identity of a talker is unimportant for speech perception under normal conditions. In fact, there are a number of studies which have provided data that are difficult to interpret in terms of independent processing of linguistic and indexical properties of speech. For example, Nygaard, Sommers, and Pisoni (1994) trained listeners over a nine-day period to identify a set of previously unfamiliar voices and associate each one with a name. After this study phase, the participants were presented with new sets of words, mixed in noise at four different levels, in a word identification task. For one group, those words were spoken by the talkers they had been familiarised with at study; for a second group the talkers were unfamiliar. Nygaard et al. found that listeners performed significantly better across all signal-to-noise ratios when they were familiar with the talkers than they did when the same words were spoken by unfamiliar talkers. Two further control groups which did not participate in the study phase were tested on the stimulus sets that were used for the trained groups; their performance was equivalent to each other and to the trained group that listened to unfamiliar voices at test. Thus, the one group which had been familiarised with the voices they heard at test clearly showed an advantage over the three other groups. These findings led Nygaard et al. to conclude that exposure to a talker's voice facilitates later recognition of new words uttered by the same talker and that, therefore, talker-specific information about voice must have been encoded in some kind of memory to be used later for recognition of novel words. Pisoni (1997) suggests that the neural representations for spoken words must include both a symbolic phonetic description and additional information about idiosyncratic characteristics of particular talkers' voices, and hence that indexical and linguistic properties of the speech signal are very closely interrelated and not independent.

In studies on word or phoneme identification in which listeners had not been familiarised with the voices they heard at test, lists spoken by multiple talkers have been shown to produce increased latencies and error rates when compared to runs in which all items are uttered by the same talker (Mullennix, Pisoni, & Martin, 1989; Nusbaum & Morin, 1992). One interpretation of this effect is that listeners in the multiple-talker condition have to make perceptual adjustments to various voices, and that this makes a call upon processing capacity. Compensating for changes in talkers thus seems to slow phonetic processing. Pisoni and Lively (1995) suggest that, by being exposed to a talker's voice, perceptual knowledge is obtained and retained in procedural memory, which might enhance processing efficiency when novel words by this talker are heard, as an analysis of idiosyncratic voice properties would not have to be carried out over and over again. They report findings from a series of experiments in which native Japanese speakers were trained to learn the English [r]/[l] contrast. Participants were given a two-alternative identification task over a 15-day training period, where [r] and [l] tokens from various environments had either been recorded from just one talker or from multiple talkers, and subsequent tests were conducted with novel words spoken by either familiar or unfamiliar talkers. The main findings were that talker variability at training facilitated robust generalisation of the newly learned phonetic contrast to new talkers; this multiple-talker advantage was still present at a follow-up study three months after the training. These results were also consistent with Nygaard et al.'s (1994) study in that familiarity with the voice(s) used at test led to enhanced performance.

In a developmental investigation, Houston and Jusczyk (2000) found that an effect of talker variability can already be detected in infants' word recognition ability. Infants at 7.5 months of age, who had been familiarised with words uttered by a female speaker, in a later test phase only responded to those words if they were also produced by another female talker, not if they were produced by a male talker. In contrast, 10.5 month olds showed evidence of generalising their knowledge of familiarised words to talkers of the opposite sex, which suggests that the ability to deal with talker variability develops over a fairly short period in the course of language acquisition.

There is thus strong evidence suggesting that talker-specific information does play a role in speech perception. Familiarity with a talker's voice has been found to improve performance on word and phoneme identification tasks, whereas performance declines in experiments which require processing of multiple unfamiliar voices as against a single unfamiliar voice. When listeners have to learn a non-native phonetic contrast, discrimination performance benefits in the long run from talker variability in the training materials. Findings like these have led researchers to propose that speech perception should be viewed in a way that is radically different from the normalisation assumption. Goldinger (1996, 1997, 1998), for example, suggests that the lexicon consists of a large number of specific instances of words which, among other attributes, include information about the voice of the talker. The listener could then use these representations to compare incoming perceptual information in an analogical rather than analytic way. In such an episodic lexicon, memory traces for words would be complex and detailed, and therefore normalisation procedures would be redundant. This perspective on word recognition has also been implemented in other models (Klatt, 1979; Johnson, 1997), which generally work on the basis of finding a direct or close match in the lexicon for the relatively unprocessed perceptual input.

One challenge for episodic models is the neuropsychological evidence suggesting that voice information is stored independently from the word recognition system (although these networks might be distributed and interconnected). Furthermore, voice information need not consist of a large collection of specific instances but might be represented abstractly. This might also be a more parsimonious model which eschews the 'head-filling-up problem', that is, the requirement for massive memory capacity which purely episodic models inevitably have (see Johnson, 1997, for a discussion). But most importantly, if talker-specific information is used in word recognition, it might be at an earlier level of processing than the lexical level; specifically, it might affect the processing of a relatively small and finite set of prelexical perceptual units. Such an influence of talker identity (and other contextual information) at a prelexical level of processing has the important advantage that a talker idiosyncrasy which affects, for example, a single phoneme contrast, can, once adjusted for by the perceptual system, generalise and thereby benefit the recognition of any word in the mental lexicon.

This potential for generalisation of prelexical representations is an essential part of many current non-episodic models of word recognition (McClelland & Elman, 1986; Norris, 1994; Stevens, 2002; Gaskell & Marslen-Wilson, 1997). There is an unresolved debate on how best to characterise the nature of prelexical representations (e.g., feature, phoneme, syllable, diphone, etc.), but these models agree insofar as that there is a level of processing mediating between early non-specific acoustic-phonetic analysis and lexical access, and that lexical representations are abstract. They also acknowledge a hierarchical structure of language processing in their architecture. A critical difference however is the extent to which levels of processing operate independently of each other, and the degree to which there is interactivity between different levels. This issue is discussed in more detail in the following section.

1.2 Feedback in models of word recognition

A central issue of debate in models of spoken word recognition concerns the question of modularity vs. interactivity. In modular or autonomous models, information in the speech signal is passed on in a bottom-up fashion to successively higher and more abstract levels of representation. In interactive models, in contrast, information flow is not just bottom-up, but higher stages in the process can pass information to lower levels and influence their behaviour. Two current models of spoken word recognition featuring an autonomous and an interactive architecture, respectively, are Shortlist (Norris, 1994), and TRACE (McClelland & Elman, 1986). Shortlist has an input phoneme layer and an output word layer, and does not permit top-down flow of information from lexical to phoneme levels. The TRACE model, on the other hand, consists of three layers corresponding to features, phonemes, and words; with bidirectional excitatory connections between levels and inhibitory connections within levels.

In the context of the debate on whether feedback should exist in models of spoken word recognition, Norris, McQueen, and Cutler (2000) have shown that a large body of experimental data can be accounted for by a processing model that is strictly bottom-up. Moreover, they reject the need for feedback on theoretical grounds, since in a functioning word recognition system, lexical feedback could do no more than confirm phonemic decisions that have already been made at the phoneme level. What feedback could do potentially is improve word recognition if the input is degraded or incomplete, by passing missing information from the lexical to the phoneme level. However, Norris et al. point out that it is debatable whether such a mechanism would be beneficial, since activation from the lexical level could also overwrite information coming from the input. Because in TRACE, for instance, phoneme nodes cannot differentiate whether activation came from the signal or from the lexical level, feedback in such a situation could lead to the system perceiving sounds that were not actually supported by the acoustics.

McClelland and Elman (1986) quote two major reasons for incorporating feedback mechanisms into the TRACE model — one is to simplify the phonemic decision making mechanism as it is integrated directly into the perceptual process; the other is to provide an integrated account of perceptual learning. By allowing the network to update itself when lexical access has been successful, learning takes place when the connections between two units that were activated simultaneously are strengthened. However, because of TRACE's architecture in which the entire system of units and connections is duplicated many times over successive time slices to account for time-invariant recognition, retuning of a connection between two units after simultaneous activation only affects this specific part of the network (time slice) and consequently does not generalise to other units (in other time slices), even if they represent exactly the same word, phoneme, or feature. The usefulness of such a learning mechanism, as implemented in the model, therefore remains questionable. McClelland and Elman's other reason for including feedback, to model phonemic decision making, has also not gone unchallenged. The Merge model (Norris et al., 2000) can accommodate experimental data from phonemic decision making, but, like Shortlist, employs only bottom-up flow of information. In Merge, phonemic decisions are made in an additional layer of nodes which receive input from both the lexical and the phoneme level, with lateral inhibition among the decision nodes. In short, Norris et al. argued that, since feedback is not necessary for explaining the available data, but disadvantageous under certain circumstances, it should not be included in models of spoken word recognition.

There are, however, data that are more difficult to explain without a feedback mechanism, although these do not necessarily require *on-line* interactivity of the type implemented in TRACE. Samuel (1997, 2001) has reported two series of experiments designed to investigate top-down lexical influence on phonemic perception, using a task that does not involve phonetic decisions. In the first study (Samuel, 1997), listeners were presented with polysyllabic words which contained [b] or [d] in the third syllable (e.g. 'inhibition', 'armadillo'). These stops had been removed and, in experimental items, replaced by signal-correlated noise, but in control items replaced by silence. When listeners hear noise instead of the stop consonant, phonemic restoration occurs, that is, the word is perceived as if the original stop was present in the signal; but the effect does not occur in the condition in which the consonant is replaced by silence. One group listened to words originally containing [b], a second group listened to items containing [d]. Before and after having listened to these stimuli, the participants were asked to categorise sounds on a [bɪ]–[dɪ] series. Samuel found a selective adaptation effect (Eimas & Corbit, 1973): The group which had listened to words in which [b] had been replaced by noise categorised fewer syllables on the continuum as [bɪ] when compared to their baseline measure taken before the experiment. The group which had listened to the [d]-items gave fewer [dɪ] responses. The control items with the silent gaps produced no such effect. These results were interpreted as evidence for top-down feedback from the lexical to the phoneme level — listeners compensated for the missing information in the signal with activation that was passed down from the lexical level; this repeated activation of either the [b] or the [d] nodes then led to the adaptation effect observed afterwards in the categorisation task, as if these sounds really had been present in the signal.

However, as the selective adaptation effect is not an on-line measure of interaction between phoneme and lexical level (the categorisation task was given to subjects only after the adaptation phase), the effect observed in

Samuel's experiment is not direct evidence of a top-down lexical effect in real-time speech processing. An alternative interpretation is that a perceptual learning mechanism operated during the adaptation phase, which modified prelexical phoneme representations over time (Norris, McQueen, & Cutler, 2003). This process would be quite different from the immediate and facilitative kind of feedback employed for instance in TRACE. If listeners in the adaptation phase learned to interpret the signal-correlated noise versions of [b] and [d] as acceptable instances of sounds belonging to those categories, the categories would be expanded and act as adaptors, which would also be consistent with the shift in categorisation on the [bɪ] – [dɪ] continuum.

In another series of experiments, Samuel (2001) used a modified procedure with [s]- and [ʃ]- final words in which the final fricative was replaced with an ambiguous sound from an [s]–[ʃ] continuum. Participants simply listened to either [s]- or [ʃ]-final items repeatedly in several blocks, which were immediately followed by a categorisation task with the [s]–[ʃ] continuum. A control group listened to items in which the last fricative had been deleted and replaced by silence. As in the previous experiment, Samuel observed a shift in categorisation on the continuum. Listeners who were exposed to [s]-final items labelled fewer sounds on the continuum as [s], and the opposite effect was obtained for listening to [ʃ]-final items. Again, no effect was found in the control group. A further experiment, in which potential cues for place of articulation in the final vowel were controlled for, produced the same pattern of results. Again, Samuel interpreted these findings in terms of selective adaptation as a consequence of lexical influence on the phoneme level; but again the possibility that listeners had learned over time to treat the ambiguous sound as an acceptable token of either of the endpoints also applies.

A study by Vroomen, van Linden, and Bertelson (2004) has directly shown that such a learning effect can occur in this situation, although in their experiments, a modulation of phonetic perception was visually-guided and did not involve lexical knowledge. They used a modified version of an experiment by Bertelson, Vroomen, and de Gelder (2003), who had demonstrated that repeated exposure to a situation which produces the well-known McGurk effect (i.e., altered phonetic perception driven by incongruent visual information of an articulating face; McGurk & MacDonald, 1976) produces visually-driven learning which can be measured after exposure for unimodally presented speech. The critical conditions in Vroomen et al.'s study involved the presentation of ambiguous tokens from an [aba]–[ada] continuum, synchronised with a video of a person who produced either [aba] or [ada]. Blocks of these exposure trials alternated with blocks of categorisation of the phonetic continuum. As in the original Bertelson et al. experiment, a shift in phonetic categorisation was observed in the test blocks such that participants were more likely to label test sounds in the way that matched the visually presented articulation with which they had previously encountered these sounds. However, while this effect occurred rapidly and reached its peak after approximately ten exposures to an ambiguous sound paired with incongruent visual information, it declined again over the course of the experiment with further exposures before finally reversing after roughly 100 repetitions of the audiovisual stimulus. These results are thus consistent with an interpretation of the Samuel (1997, 2001) experiments (which also consisted of hundreds of adaptation trials) by which lexically-driven perceptual learning occurred during the first few trials, and the newly-adjusted categories then acted as adaptors for the remainder of the experiments.

To investigate the perceptual learning account in a more direct way, Norris et al. (2003) conducted an experiment in which participants listened to words ending in [f] or [s]. For one group of subjects, the final [f] sounds in these words were replaced with an ambiguous fricative midway between [f] and [s], but the [s]-final words remained natural. A second group received words manipulated in the reverse pattern, with natural sounding [f]-final words and the ambiguous fricative [?] replacing [s] sounds. A control group listened to a set of nonwords which also had the ambiguous sound in final position. For all three groups, these items were presented interspersed with other words and nonwords that contained neither [f] nor [s] in the context of a lexical decision task, which served as the exposure phase of the experiment. Overall, in the experimental groups 90% of [?]-final items were accepted as real words. After the exposure phase, all participants were asked to categorise sounds from a five-step [ɛf]–[ɛs] continuum (the same series from which the ambiguous [?] had been selected). Results showed that, when compared to the control group, participants who had listened to the natural [s]-final words and ambiguous [f]-final words were more likely to categorise sounds on the continuum as [f], whereas those who had received the reversed training categorised more sounds as [s].

In the context of the feedback debate, these findings thus provide support for a rapid perceptual learning mechanism. As the effect only occurred when the ambiguous fricative sound was presented in words, but not in the context of nonwords, there is also direct evidence for flow of information from lexical to prelexical levels of processing, and the same conclusion follows from either interpretation of the Samuel (1997, 2001) experiments. However, contrary to the interpretation given by Samuel, this kind of lexical feedback is seen by Norris et al. as a mechanism that works off-line, and over a longer period of time than the instantaneous feedback employed in interactive models of speech perception. Off-line lexical learning also may be driving other apparent on-line effects (McQueen, 2003). In particular, recent studies by Samuel and Pitt (2003) and Magnuson, McMurray, Tanenhaus, and Aslin (2003) have been interpreted as evidence that prelexical compensation for coarticulation can be affected by lexical feedback. In their experiments, an ambiguous fricative sound at the end of a word or pseudoword differentially affected listeners' reports of a following word-initial stop consonant, depending on on the lexical status of the first item, which the authors explained in terms of an on-line lexical bias on the ambiguous fricative. Again there are alternative explanations based on learning, however, leaving the issue of on-line feedback in speech recognition open for further investigation (see McQueen, 2003, for discussion). Learning may be both of a stochastic nature and operate over long periods of time (i.e., the probability of a particular fricative–stop sequence, as well as the surrounding phoneme context, varies in natural language and thus has effects on prelexical processing), or of the relatively short-term type that has been found in the experiments by Norris et al. (2003).

A perceptual learning mechanism which is driven by stored lexical knowledge and which affects the mapping of acoustic cues to prelexical perceptual units over time is extremely interesting in the context of the existing literature on perceptual adjustments in speech processing. There are abundant findings in the literature which suggest that listeners learn to adapt to sources of variability in the speech signal such as talker idiosyncrasies, dialects, foreign accents, or artificial signal manipulations. There are proposals for explicit mechanisms (e.g., stochastic models; Maye, Werker, & Gerken, 2002), which can explain bottom-up learning over long periods of time, such as acquiring the phoneme inventory or the phonotactics of a native language. The lexically-driven perceptual learning reported by Norris et al., in contrast, provides an account of short-term modulations in the speech perception system that result from disambiguating difficult speech input. The largest part of this thesis aims to characterise the nature of this type of perceptual learning better, and to establish some of the constraints under which it operates.

1.3 Perceptual Learning

The role of plasticity in the human brain is to enable its systems to adapt to environmental factors that either cannot be anticipated by genetic programming, or would require too much of it (Rauschecker, 1999). Such environmental factors may be trauma, in an extreme case, but adaptation can also occur as a consequence of mere exposure and learning, and is an ability which infants already have (e.g., Cheour et al., 2002; Chollet, 2000; Morris, 1997). In auditory perception, there is ample evidence showing that exposure or training on auditory stimuli can result in detectable changes in underlying neurophysiological processes (e.g., Kraus et al., 1995; Kraus, 1999; Rauschecker, 1999; Titova & Näätänen, 2001; Tremblay, Kraus, Carrell, & McGee, 1997; Jacquemot, Pallier, LeBihan, Dehaene, & Dupoux, 2003). Chapter 4 investigates the neural systems in auditory cortex that may be implicated in the type of plasticity that occurs as a result of lexically-driven learning.

Neural mechanisms become generally less plastic with age, so that recovery from trauma, and learning in certain domains, becomes increasingly difficult. For speech perception, and in particular the formation of phonetic categories, this process appears to occur already at a surprisingly early age. Pallier, Bosch, and Sebastián-Gallés (1997) tested bilinguals, who had learned their second language (Spanish and Catalan) before the age of six, on their ability to categorise and discriminate a vowel height contrast which appears in Catalan but not in Spanish, using a synthesised 7-step [e]–[ɛ] continuum. Half of the participants had Spanish-speaking parents, the other half had Catalan-speaking parents. While Catalan-born listeners labelled the two vowels categorically and exhibited heightened discrimination performance, Spanish-born listeners showed none of these effects. A further experiment in which participants were asked to rate the goodness of each of the seven stimuli relative to one Spanish word containing [e] and two Catalan words containing [e] and [ɛ], largely confirmed the results of the first experiment. Spanish-born bilinguals only exhibited a slight trend in preference when rating the two vowels, implying at least some awareness of the non-native contrast, but Catalan-born subjects showed a clear discrimination. This suggests that it is very difficult to learn non-native phonetic contrasts even at a very early age and after extensive exposure. Late learners, however, have been shown to be still less sensitive to unfamiliar language-specific contrasts than early learners (MacKay, Flege, Piske, & Schirru, 2001), suggesting that plasticity in the auditory system, as in other cognitive domains, declines with age.

Phonetic categories appear to be established in the first few months of life, and become less flexible with age. When phonetic categories are in place, they affect later perception (Werker & Tees, 1984). Kuhl and Iverson (1995) propose that there is a general mechanism by which language experience alters phonetic perception in a way that distorts perceived distances between speech sounds. For both adults and infants, the difference between the prototype of a sound and a sound close to it is not perceived as easily as the difference between a non-prototypical sound and a sound close to it, even when the acoustic distance within the two pairs is equal (Liberman, Harris, Hoffman, & Griffith, 1957). Cross-linguistic research on US-American and Swedish six month-olds showed that exposure to their native language had altered their phonetic perception, as American infants exhibited the categorical perception effect for the native vowel [i] but less so for the Swedish rounded [y], whereas for Swedish infants the reverse pattern emerged (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). The authors concluded that mere listening to a language affects perception of the phonetic units of all languages.

Adults do not lose entirely the ability to learn new phonetic contrasts, however. Tremblay et al. (1997) trained native English speakers to identify synthetically generated labial stops on a continuum varying in VOT. Their listeners' task was to label two prevoiced sounds as either [mba] or [ba]. After a five-day training period, participants showed better performance on discrimination and identification of the previously unfamiliar contrast when compared to their baseline measure and a control group. Interestingly, it was also found that the improvement generalised to a new continuum which differed from the one used in the training in place of articulation ([nda] – [da]). Furthermore, the training effect was reflected in electrophysiological measurements in that the Mismatch Negativity (MMN) increased in duration and spatial extent, and that the onset of the mismatch response decreased as a result of training. The findings of Tremblay et al. demonstrate that the adult perceptual system is still plastic and capable of accommodating new categories after training. However, what their study, and others, also has shown is that this usually requires explicit training on the order of weeks or even months (Tremblay et al., 1997; Kraus et al., 1995; Lively, Logan, & Pisoni, 1993; Pisoni, Aslin, Perey, & Hennessy, 1982), contrary to the rapid effects in the studies by Samuel (2001) and Norris et al. (2003). Modification of an existing native category appears to require much less exposure and no explicit training, and even occurs without listeners being aware of the change.

Acquiring the phonetic categories of a native language during the first years of life has been described as a warping of the perceptual space, which has the consequence that sounds that fall inside these categories are recognised more readily and reliably, whereas sounds that fall outside a category can be discriminated from relevant speech sounds more easily. Gibson (1969) stated that

> “Perceptual learning then refers to an increase in the ability to extract information from the environment, as a result of experience and practice with stimulation coming from it. That the change should be in the direction of getting better is a reasonable expectation, since man has evolved in the world and constantly interacts with it. Adaptive modification should result in better correlation with the events and objects that are the sources of stimulation as well as an increase in the capacity to utilize potential stimulation” (pp 3–4).

Following Gibson’s definition, perceptual learning in speech, as in other domains, has an adaptive function, and its purpose in speech is to aid future comprehension. The role for short-term adjustments to a phonetic category boundary based on lexical knowledge as observed by Norris et al. (2003) is then to compensate for those types of variability that naturally occur within a language rather than between languages. Candidate sources of such variability have already been mentioned above, and include in particular inter-talker variation (e.g., vocal tract characteristics, dialects, or speech impediments) and intra-talker variation (e.g., affect, register, or speaking rate). The experiments in Chapter 2 examine the compensatory role of learning by testing whether adjustments are talker specific. Specifically, is a modulation of the category boundary induced in this way applied only to that talker whose speech triggered the adjustment in the first place? Or, is the modulation applied broadly, such that it affects any similar sounds on future encounters regardless of who produced them? The former outcome in particular would have implications for models of spoken word recognition, as these would need to include a way of maintaining talker-specific representations. The experiments further addressed a second aspect of the specificity of learning by asking whether the adjustment to the category boundary hinges on acoustic cues that are intrinsic to the category, or, alternatively, whether other cues that are external to the category have an effect, that is, whether the context in which the critical sounds occur is important.

1.4 Neural mechanisms

A description of the neural mechanisms that may be involved in perceptual learning requires an understanding of the computational processes that accomplish a change in the relationship between successive levels of processing. Furthermore, the degree to which such processes are automatic or require attentional resources, as well as the processes that achieve consolidation of learning, need to be established. In this section, current findings and theories from speech perception and other domains on each of these three issues will be addressed in turn.

How might a lexically-driven perceptual adjustment of prelexical processing be conceptualised in terms of a computational learning mechanism? More than 50 years ago, Donald Hebb proposed a learning mechanism at the neuronal level which could explain how connections between individual cells are strengthened in an unsupervised and adaptive fashion. Hebb's original theorem stated that "when an axon of cell A is near enough to excite cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1949, p 62). This principle is now well established in neuroscience, with an important more recent addition which qualifies "near enough" and also accounts for a weakening between connections (Sejnowski, 1999). The crucial factor which decides whether a connection is made more or less efficient is timing: if the presynaptic cell fires immediately before the postsynaptic cell, the connection between them becomes more efficient (long-term potentiation); if the presynaptic spike occurs spuriously, or just after the postsynaptic cell has discharged, it is unlikely to have participated in the postsynaptic firing and the connection between the cells is therefore weakened (long-term depression; Rao & Sejnowski, 2001; Stuart & Sakman, 1994).

The principle of Hebbian learning has also been widely applied at a systems level, both in neurobiology and in cognitive psychology. At the level of neuronal populations, learning may have a variety of consequences on the interplay between neural networks both spatially and temporally (Gilbert, Sigman, & Crist, 2001; Sterr, Elbert, & Rockstroh, 2002). From a cognitive perspective, Hebbian learning is a parsimonious account of unsupervised change in the relationship of two levels of processing in a hierarchical system with only feedforward flow of information. The limitation of the Hebbian principle is, however, that it is by itself not applicable when learning spans more than two levels of processing. In the context of a lexically-driven adjustment to a phonetic category boundary, the mechanism can only work with additional feedback in the system. Consider the situation of the Norris et al. (2003) experiment, where, for example, an acoustic pattern is consistent to an equal extent with the prelexical phonetic representations for [f] and [s], and occurs in a sequence where it is lexically consistent only if it were an [f]. The ability of the [f] category to activate the lexical item would be strengthened through Hebbian learning. The locus of the perceptual learning, however, is at the level of the mapping from the acoustic pattern to a category, and this mapping would remain unaffected: A subsequent encounter with the ambiguous sound would still activate both categories equally. The data reported by Norris et al. therefore suggest the existence of a feedback mechanism built into the perceptual system, by which prelexical processing is altered according to a training signal which originates from the lexical level of processing. This can in principle be possible with the type of on-line feedback that is implemented in the TRACE model (recall though, that because of TRACE's architecture, learning can not generalise across the network). As was noted earlier, Norris et al. have proposed that perceptual learning is driven by a different type of feedback, which operates off-line, and alters the mapping from acoustic cues to prelexical representations over a longer period of time. In the context of their study, this top-down flow of information could then facilitate the expansion of the [f] category in the case of [f]-biased exposure, or of the [s] category in the case of [s]-biased exposure.

The question arises then whether this modification of phonetic categories is induced in an automatic and preattentive fashion, or whether attentional resources are required to instigate such change. Pylyshyn (1999) has argued that, in the case of perceptual learning in the early visual system, the role of attention may be to introduce a bias according to which only relevant properties of the input are selected for further processing. Similarly, Raichle (2001) suggested that higher cognitive processes may be able to affect the output of early perceptual modules when a perceptual analysis has been completed by selecting among several possibilities, and thus having the ability to act like filters in case of ambiguous output. Both authors generally reject the notion of cognitive penetrability, by which there would a direct and immediate influence of higher cognitive processes on earlier stages; instead they argue for what Raichle has termed 'off-line penetrability', a perceptual learning mechanism which may in the long run use information from high levels of processing and is mediated by attention; this view is thus very similar to what Norris et al. (2003) suggested for off-line perceptual learning. Applied to speech perception in the case of ambiguous input in the acoustic signal, off-line penetrability would then refer to the listener's usage of lexical or semantic knowledge to infer the identity of the sound the talker intended to produce. Attention would bias the interpretation of the output of an early perceptual stage, so that only informative attributes are further processed. When this happens repeatedly, attention may become redundant as processing becomes automatic (Goldstone, 1998).

Certain types of learning and memory require a period of consolidation to become effective, and there is a growing body of evidence from different cognitive domains that sleep, in particular in the rapid-eye-movement (REM) stages, is essential for this process to occur (Hobson & Pace-Schott, 2002; Sejnowski & Destexhe, 2000; Walker & Stickgold, 2004; Walker, Brakefield, Hobson, & Stickgold, 2003). The main lines of evidence are that REM phases are prolonged after learning, and that disruption of REM sleep appears to have negative consequences on learning. Others dispute this hypothesis, mainly on the grounds that disruption of sleep causes stress and that for this reason many observed disadvantages for learning and memory consolidation might be artefactual (e.g., Siegel, 2001; Vertes & Eastman, 2000). Without making any claims about which sleep stages are implicated, a number of studies have now shown increased performance on procedural learning tasks after a period of natural sleep as compared to natural waking. In the auditory domain, for instance, it has been found that subjects perform better on a pitch discrimination task after they had a night's sleep than after an equivalent time interval of waking (Gaab, Paetzold, Becker, Walker, & Schlaug, 2004). Gottselig, Hofer-Tinguely, Borbély, Rétey, and Achermann (2004) showed a benefit both of sleep and restful waking over busy waking for performance on a task that required the learning of a complex tone sequence. Sleep has also been shown to play a role in consolidation processes that are related to declarative memory, as in the acquisition of novel words (Gaskell & Dumay, 2005).

There is some recent evidence suggesting that the speech perception system can benefit from sleep when it has to adjust to difficult listening conditions. Fenn, Nusbaum, and Margoliash (2003) measured listeners' performance on transcribing synthetic speech, which is relatively hard to understand without any training (∼30% correct transcriptions). After a training procedure involving explicit feedback, participants' performance increased significantly by ∼25% when tested immediately afterwards with new synthesised speech materials. Relative to this gain, participants who were tested after a period of 12 hours spent awake showed a significant drop in performance, while listeners who had slept during a 12-hour interval performed equally well as those who were tested immediately after training. Fenn et al. suggested that this finding demonstrates sleep-dependent consolidation of procedural learning in the mapping of an acoustic pattern to linguistic categories, which furthermore showed generalisation to novel test items.

Some questions regarding the neural mechanisms that are responsible for perceptual learning in speech are addressed in chapters 3 and 4. The focus there will be on the role of sleep and the neural substrates of learning. Specifically, the experiment in chapter 3 tests whether a perceptual adjustment to the speech of a particular talker remains stable over time. Two conditions are compared, where in one condition participants have the opportunity to consolidate learning during sleep after perceptual learning has occurred, and in a second condition participants are tested after an equivalent interval of waking. The experiment also addresses whether an adjustment is affected by hearing unambiguous tokens of the critical sounds that were produced by talkers other than the exposure talker.

A necessary preliminary to understanding the neural basis of perceptual adjustments of the prelexical processing system is to identify the cortical regions that are implicated. In chapter 4, an experiment is presented which uses functional magnetic resonance imaging to investigate how prelexical processing of speech, and perceptual learning within this system, may be instantiated in the neuroanatomy of the human brain.

References

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

Cheour, M., Martynova, O., Näätänen, R., Erkkola, R., Sillanpää, M., Kero, P., et al. (2002). Speech sounds learned by sleeping newborns. *Nature*, *415*, 599–600.

Chollet, F. (2000). Plasticity in the adult human brain. In A. W. Toga & J. C. Mazziotta (Eds.), *Brain mapping: The systems.* San Diego, CA: Academic Press.

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

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*, 614–616.

Gaab, N., Paetzold, M., Becker, M., Walker, M. P., & Schlaug, G. (2004). The influence of sleep on auditory learning: A behavioral study. *Neuroreport*, *15*(5), 731–734.

Gaskell, M. G., & Dumay, N. (2005, July). Plasciticy in lexical competition: The impact of vocabulary acquisition. In V. Hazan & P. Iverson (Eds.), *Proceedings of the ISCA workshop on plasticity in speech perception* (p. 108).

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* Englewood Cliffs, NJ: Prentice-Hall.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron*, *31*, 681–697.

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*, 1166–1183.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech perception.* San Diego, CA: Academic Press.

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

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Gottselig, J. M., Hofer-Tinguely, G. H., Borbély, S. J., Rétey, J. V., & Achermann, P. (2004). Sleep and rest facilitate auditory learning. *Neuroscience*, *127*, 557–561.

Halle, M. (1985). Speculations about the representation of words in memory. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honour of Peter Ladefoged.* Orlando, FL: Academic Press.

Hebb, D. O. (1949). *Organization of behavior: A neuropsychological theory.* New York: John Wiley and Sons.

Hobson, J. A., & Pace-Schott, E. F. (2002). The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning. *Nature Reviews Neuroscience*, *3*, 679–693.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, *23*, 9541–9546.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Kraus, N. (1999). Speech sound perception, neurophysiology, and plasticity. *Journal of Pediatric Otorhinolaryngology*, *47*, 123–129.

Kraus, N., McGee, T., Carrell, T. D., King, C., Tremblay, K., & Nicol, T. (1995). Central auditory system plasticity associated with speech discrimination training. *Journal of Cognitive Neuroscience*, *7*(1), 25–33.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the perceptual magnet effect. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research.* Baltimore: York Press.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

Liberman, A. M., Harris, K. S., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.

Lively, S., Logan, J., & Pisoni, D. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255.

MacKay, I. R. A., Flege, J. E., Piske, T., & Schirru, C. (2001). Category restructuring during second-language speech acquisition. *Journal of the Acoustical Society of America*, *110*(1), 516–528.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science*, *27*, 285–298.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

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

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

McQueen, J. M. (2003). The ghost of Christmas future: Didn't Scrooge learn to be good? Commentary on Magnuson, McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science*, *27*, 795–799.

Morris, R. G. M. (1997). Learning, memory and synaptic plasticity: Cellular mechanisms, network architecture and the recording of attended experience. In D. Magnuson (Ed.), *The lifespan development of individuals: Behavioral, neurobiological, and psychosocial perspectives.* New York, NJ: Cambridge University Press.

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*, 365–378.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299–370.

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

Nusbaum, H., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: Ohmsha.

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

Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*, B9–B17.

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., et al. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain*, *117*, 1283–1301.

Pisoni, D. B. (1997). Some thoughts on 'normalization' in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–30). San Diego, CA: Academic Press.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 297–314.

Pisoni, D. B., & Lively, S. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research.* Baltimore: York Press.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–423.

Raichle, M. E. (2001). Bold insights. *Nature*, *412*, 128–130.

Rao, R. P. N., & Sejnowski, T. J. (2001). Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Computation*, *13*, 2221–2237.

Rauschecker, J. P. (1999). Auditory cortical plasticity: A comparison with other sensory systems. *Trends in Neurosciences*, *22*, 74–80.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical perception* (pp. 89–112). Cambridge, UK: Cambridge University Press.

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

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

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

Sejnowski, T. J. (1999). The book of Hebb. *Neuron*, *24*, 773–776.

Sejnowski, T. J., & Destexhe, A. (2000). Why do we sleep? *Brain Research*, *886*, 208–223.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech perception with primarily temporal cues. *Science*, *270*, 303-304.

Siegel, J. M. (2001). The REM sleep–memory consolidation hypothesis. *Science*, *294*, 1058–1063.

Sterr, A., Elbert, T., & Rockstroh, B. (2002). Functional reorganization of human cerebral cortex and its perceptual concomitants. In M. Fahle & T. Poggio (Eds.), *Perceptual learning.* Cambridge, Ma.: MIT Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

Stuart, G. J., & Sakman, B. (1994). Active propagation of somatic action potentials into neocortical pyramidal cell dendrites. *Nature*, *367*, 69–72.

Titova, N., & Näätänen, R. (2001). Preattentive voice discrimination by the human brain as indexed by the mismatch negativity. *Neuroscience Letters*, *308*(1), 63–65.

Tremblay, K., Kraus, N., Carrell, T. D., & McGee, T. (1997). Central auditory system plasticity: Generalisation to novel stimuli following listening training. *Journal of the Acoustical Society of America*, *102*(6), 3762–3773.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex*, *24*, 195–209.

Van Lancker, D. R., Kreiman, J., & Cummings, J. L. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology*, *11*(5), 665–674.

Vertes, R. P., & Eastman, K. E. (2000). The case against memory consolidation in REM sleep. *Behavioral and Brain Sciences*, *23*, 867–876.

Vroomen, J., van Linden, S., & Bertelson, P. (2004). *Recalibration and selective adaptation of ambiguous speech sounds co-occur.* Paper presented at the 45th Meeting of the Psychonomic Society, Minneapolis, Minnesota, November 18 – 21.

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*, 616–620.

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron*, *40*, 121–133.

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.

Chapter 2

Specificity of Perceptual Learning

A version of this chapter has appeared as F. Eisner & J.M. McQueen (2005) The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*, 224–238.

2.1 Introduction

This series of experiments is concerned with the nature of perceptual adjustments that take place in the speech recognition system in response to unusual speech production. The process of decoding speech is necessarily complex, to a great extent due to the fact that, in addition to structural variation such as coarticulation, speech is characterized by a large amount of both inter- and intra-talker variability. The realization of a given phoneme varies within individuals as a function of, for example, voice quality, emotional state, or speaking rate. Inter-individual differences, the focus of the present series of experiments, are caused by factors such as vocal tract shape, accent, or articulatory habits (see, e.g., Klatt, 1986, 1989). The cumulative effect of all these sources of variability is that the mapping from input to categories is a many-to-many problem: Not only can one phoneme have different acoustic realizations, but one acoustic pattern can elicit different phonemic percepts (Nusbaum & Magnuson, 1997; Repp & Liberman, 1987). How do listeners deal with this variability? A number of previous studies have shown that the perceptual system dynamically adjusts to speech that is initially difficult to understand. The characteristics of the mechanism that achieves perceptual constancy and the constraints under which it operates are largely unknown, however. In this study we asked whether there are talker-specific adjustments in the speech perception system in response to unusual productions of speech sounds, and if so, how detailed those adjustments are.

Evidence of such dynamic adjustments, as indexed by improved intelligibility after sufficient exposure, has been found with synthetic speech (Greenspan, Nusbaum, & Pisoni, 1988; see also Maye, Aslin, & Tanenhaus, 2003), noise-vocoded speech (Hervais-Adelman, Johnsrude, Davis, & Brent, 2002), and compressed speech (Dupoux & Green, 1997; Mehler et al., 1993). Such studies have revealed important constraints on perceptual learning in speech processing. Greenspan et al., for example, note that variability in the training materials is crucial for learning; repetitions of a small set of stimuli did not produce improved intelligibility in their study. Hervais-Adelman et al. observed that adaptation to noise-vocoded speech was absent when listeners were presented with phonotactically legal nonword sentences. This suggests that higher-level (e.g., lexical) information is required for adaptation to occur.

Perceptual learning has also been shown in response to natural but accented speech input. Moving to a different dialectal environment often requires adaptation to unfamiliar input from a whole community of talkers. British English speakers who have lived in the United States, for example, learn to recognize an alveolar tap [ɾ] as an instance of [t] (Scott & Cutler, 1984). Similarly, American immigrants to Britain may have to learn that the glottal stop [ʔ] is an instance of the same phoneme. Adjustments are also made in response to talkers who speak a language with a non-native accent. Clarke (2002, 2003) observed that after short exposure to Spanish-accented American English, listeners performed faster on a task that required matching a visual stimulus to accented auditory input than control listeners who had had exposure to another voice talking in non-accented English. Bradlow and Bent (2003), in a training–test paradigm, investigated perceptual adjustment to Chinese-accented English. One group of listeners who had heard multiple talkers, and another group that had heard only the test talker at training, performed equivalently, and better than other training groups, on a transcription task. Listeners who heard only a single talker at training, one who was different from the test talker, did not show improved performance. These results suggest that perceptual adaptation is useful when the same talker is encountered again, and, furthermore, that adaptation in response to a single talker does not generalize to another talker. However, if there is variability in the input, as introduced by multiple talkers, the perceptual system appears to be able to extract abstract information about the accent that can be used to facilitate comprehension of other talkers with the same accent. Because of the nature of Bradlow and Bent's task, however, the type of information extracted (e.g., featural, segmental, prosodic, or rhythmic) cannot be determined. Nevertheless, as was shown in another study (Evans & Iverson, 2004), it is possible that learned characteristics of an accent can constrain the interpretation of subtle phonetic cues.

Some studies on accent normalization have used intelligibility of words or sentences as a dependent measure, and hence provide few cues as to the level or detail of adjustment. While others have however examined the role of phonetic detail in processing accented speech (Evans & Iverson, 2004; Scott & Cutler, 1984), adjustments in these cases were the outcome of exposure to a whole language community, possibly over many years, and are therefore not necessarily carried out by the same mechanism as that responsible for individual talker normalization. In the present study, in contrast, we sought to evaluate the degree of detail that listeners learn about the characteristics of an individual talker's speech after short-term exposure.

There is abundant evidence that perceptual adjustments are made to the speech of individual talkers. Classic studies by Ladefoged and Broadbent (1957) and Ladefoged (1989) have shown that listeners evaluate a talker's vowel space and apply this computation in interpreting following vowels within the same utterance. More recently, Nygaard, Sommers, and Pisoni (1994) found that spoken word identification was improved for listeners who had previously been familiarized with the talkers' voices, compared to control listeners who had been familiarized with another set of voices (see also Nygaard & Pisoni, 1998). Their findings suggest that once an adjustment to the idiosyncrasies of a particular talker's utterances has been made, the result of this process is stored and will be used again to facilitate perception when this voice is encountered at a later point — a conclusion that is in line with the recurrent observation that listeners encode details of talkers' voices in long-term memory (Church & Schacter, 1994; Martin, Mullennix, Pisoni, & Summers, 1989; Goldinger, 1996, 1998; Goldinger, Pisoni, & Logan, 1991; Palmeri, Goldinger, & Pisoni, 1993; Pisoni, 1993).

In an earlier study, Mullennix and Pisoni (1990) investigated directly a possible influence of talker specific information on linguistic processing. They employed a same–different classification task of either the dimension voice or the dimension phoneme (a word-initial voicing contrast). While in the experimental conditions the respective other dimension was always varied, response latencies were compared to a control condition where the other dimension was held constant. Results showed that variation of these two dimensions produced mutual interference, suggesting that they are not processed independently of each other. However, there was an asymmetry such that variation in voice caused more interference with phoneme classification than vice versa. Given this asymmetry, Mullennix and Pisoni concluded that linguistic processing is contingent on voice processing, more specifically, that talker information is extracted from the signal first and then influences phonetic processing (see also Green, Tomiak, & Kuhl, 1997; Lattner, 2002; Knösche, Lattner, Maess, Schauer, & Friederici, 2002).

Another observation on the constraints of a perceptual learning mechanism is that the initial adjustment to a talker comes at a processing cost. Mullennix, Pisoni, and Martin (1989) report that identification and naming of a list of words in noise deteriorates and slows down when these words are produced by multiple, intermixed talkers — relative to a list where all words are produced by the same talker. Mullennix et al. propose that the perceptual system must engage in an adjustment process each time a novel voice is encountered. On the other hand, when there is only one talker in the set, the system is already in the right configuration at the time a word is presented, leading to better identification performance and shorter response latencies. Nusbaum and Morin (1992) report a similar and consistent effect of multiple-talker compared to single-talker presentations in response latencies to vowels, consonants, and words.

A recent study by Norris, McQueen, and Cutler (2003) provides some insight into how a perceptual learning mechanism in speech perception might operate. This study demonstrated a lexically-driven modulation of the category boundary for a consonant contrast, which was induced in an exposure phase and measured in a subsequent phonetic categorization task. In the exposure phase listeners heard naturally produced words, some of which were edited. For one group of listeners, all instances of the fricative sound [s] were replaced by a perceptually ambiguous sound lying midway between [s] and [f]. For another group of listeners all cases of [f] were replaced by the same ambiguous fricative sound. Results showed that the group which had heard the ambiguous sound in [s]-biased lexical contexts categorized more sounds on an [f]–[s] continuum as [s], while the other group categorized most sounds as [f]. In accord with what Hervais-Adelman et al. (2002) found for noise-vocoded speech, this study thus shows that a perceptual adjustment is made when an idiosyncratic production of a speech sound is placed in an appropriate lexical context.

Previous studies have therefore shown that the speech perception system makes adjustments to both natural speech and to speech that is in some way unusual. The adjustment requires processing capacity, and evidence for one specific adjustment mechanism, which is lexically driven, has been found. Patterns extracted from these adjustments are stored and the information is re-used when speech with similar characteristics is encountered again.

A number of important questions about the constraints of perceptual learning remain unanswered, however. In the current study we addressed two issues regarding its specificity. First, it is not clear how detailed the adjustments are: Are adjustments made at a segmental level (i.e., with respect to individual phonemes), at a lexical level (i.e., with respect to individual words), or more globally (e.g., with respect to pitch characteristics of a talker's voice)? Second, it is not clear whether the effect of perceptual learning is applied talker-specifically, or whether it also affects processing of speech from other talkers. While studies on accent learning have shown that the outcome of perceptual adjustment is of benefit for comprehension when, subsequently, talkers with the same accent are encountered (Bradlow & Bent, 2003; Scott & Cutler, 1984), it is uncertain whether such learning may be misapplied to other talkers who do not have that accent. Similarly, the outcome of individual talker normalization is clearly beneficial when listening to the speech of the same talker again (Nygaard et al., 1994), but may have a detrimental effect when applied to another talker. These two issues, that of the level and detail of application of learning, and that of generalization to other talkers, were investigated using the exposure–test paradigm developed by Norris et al. (2003). We chose this paradigm because it provides tight control over the learning effect (the bias in the interpretation of an ambiguous sound is determined by lexical knowledge alone, not by differences in the ambiguous sound, or any other sounds, between conditions). It is therefore well suited to test whether the learning effect is specific to a phonetic contrast alone. Furthermore, the paradigm allows testing for talker-specificity of the adjustment. Listeners were exposed to edited, natural speech coming from one talker, and then tested for a perceptual learning effect with materials made from another talker's utterances.

2.2 Experiment 1

The aim of Experiment 1 was to examine whether perceptual learning after exposure to a (female) talker with unusual fricative productions would generalize to a test situation where listeners are presented with a new (female) talker. Conditions where the talker at test and exposure was the same (replicating the experimental conditions of Norris et al. (2003) served as a comparison for the talker-change conditions. Two further control conditions (identical to the nonword conditions of Norris et al. (2003) were included to provide a measure of the extent to which the adjustment is lexically driven.

A pretest was conducted in order to find a fricative [f]–[s] sound which was sufficiently ambiguous to Dutch listeners. The main experiment then consisted of an exposure phase (auditory lexical decision) followed by a brief test phase (phonetic categorization). There were four exposure conditions as defined by the types of words and nonwords used in the lexical decision task. In one experimental condition, all twenty instances of [s] (in word-final position) were replaced by a perceptually ambiguous sound [?], while all twenty [f] sounds (also in word-final position) remained natural. A second condition consisted of items in which all the [f] sounds were replaced by [?], but all [s]'s were natural productions. Two control groups listened to the ambiguous sound [?] in nonword contexts, where one group additionally received naturally-produced [f]-final words and the other group natural [s]-final words. As in the Norris et al. (2003) study, these groups were used to control for the possibility that an effect in the experimental conditions was due to selective adaptation or contrast effects, as opposed to a lexical effect (see Norris et al., 2003, for discussion). These four groups were then tested on an ambiguous [ɛf]–[ɛs] continuum, made from materials constructed from utterances of the talker of the exposure phase. Given that these conditions were an exact replication of the Norris et al. study, using the same words and procedure but a different talker, we expected to replicate the earlier results. The experimental exposure group which listened to ambiguous [f]-final and natural [s]-final words should subsequently categorize more sounds on the [ɛf]–[ɛs] continuum as [f], while the other experimental exposure group should categorize more sounds as [s]. The control groups should give intermediate responses and were not expected to differ from each other.

Our main interest, however, was in two further groups of participants who listened to the stimuli of the two lexically-biased exposure conditions, but were then tested on an [ɛf]–[ɛs] continuum in which the vowel [ɛ] came from an utterance by a novel talker who was also female and was similar in age to the exposure talker. Since vowels are a rich source of talker identity information, this manipulation was expected to signal a change in talkers between exposure and test. If perceptual learning generalizes to another talker, a shift in category boundary as a function of exposure condition should be evident in the categorization data. That is, the categorization data for these groups should show the same pattern as those for the listeners in the lexically-biased exposure conditions who were tested on the exposure talker. If, however, perceptual learning does not generalize to a different talker, listeners should not apply a previously learned adjustment when they notice a change in talkers. There should therefore be no difference in categorization performance between the two novel talker test groups.

2.2.1 Method

Participants

A total of 105 native speakers of Dutch drawn from the MPI for Psycholinguistics participant pool took part in the experiment. Nine volunteers participated in the pretest and the remaining 96 in the main experiment. None of them reported any hearing disorders. All were paid for their participation.

Pretest

Stimulus construction A number of tokens of the three syllables [ɛf], [ɛs], and [ɛx], produced by a female native speaker of Dutch, were recorded in a sound-damped booth onto digital audio tape (DAT). Recordings were redigitized at a 16 kHz sampling rate and 16-bit quantization on a Sun Sparc workstation and edited with Xwaves. One token each of [ɛf] and [ɛs] was selected to create an [ɛf]–[ɛs] continuum. The fricatives were excised from the vowel at a zero crossing at the onset of frication energy, and edited to match the mean duration and intensity of [f] and [s] in spoken word contexts. These mean duration and intensity values (202.4 ms and 55.2 dB SPL) were derived from measurements of the experimental items recorded for the lexical decision part of the experiment (see below). The waveforms of both fricatives were cut, then linearly smoothed at offset over a 75 ms window, and finally scaled to be of equal intensity. The resulting [s] and [f] sounds were then used to make a 21-step continuum, employing an algorithm that combined each of the two sounds sample by sample in 21 graded proportions, such that step 1 was the original [f] and step 21 the original [s], with 19 equally spaced steps in between (McQueen, 1991). Each step was then spliced onto the vowel [ɛ], which was isolated from one of the [ɛx] syllables and which was 112 ms in duration. A vowel from a velar context was used in order to avoid transitional cues to labiodental [f] or alveolar [s] place of articulation. (Note that Norris et al., 2003, used vowel tokens that always cued labiodental place, which resulted in a residual [f]-bias.)

Procedure Informal listening by four native Dutch speakers indicated that the most ambiguous range of the [ɛf]–[ɛs] continuum was steps 6–15. These ten syllables were presented to the pretest listeners over closed headphones in a sound-damped booth. Items were pseudo-randomized by concatenating ten individually randomized lists containing one of each syllable. There was a short practice sequence in which each token was played once. Responses were made by pressing one of two buttons labelled ‘F’ and ‘S’, counterbalanced for handedness across the sample such that one half of the participants made ‘F’

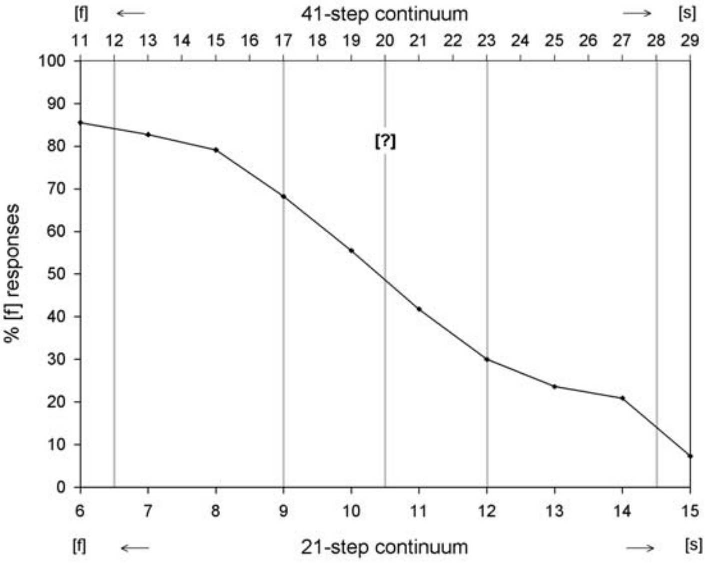

Figure 2.1: Experiment 1, pretest: Mean percentage of [f] responses to each of the ten steps. The most ambiguous point of the continuum lies between steps 10 and 11, corresponding to step 20 on a 41-step continuum (top x-axis).

responses and the other half ‘S’ responses with their dominant hand. Items were presented at a rate of 2.6 s between syllable onsets.

Results Percentages of [f] responses are plotted against each of the ten steps of the continuum in Figure 2.1. The continuum was judged by listeners to be most ambiguous (50 percent [f] responses) at the point midway between steps 10 and 11. Hence a more fine-grained 41-step continuum was made from the endpoint stimuli using the technique described above. Step 20 corresponded to step 10.5 on the 21-step continuum. This step [?] was then used to make the ambiguous items in the exposure phase of the main experiment and, along with steps 12, 17, 23, and 28 (corresponding to 85, 70, 30, and 15 percent of [f] responses, respectively), was also used in the phonetic categorization phase.

Materials and Stimulus Construction

Lexical decision Stimuli were constructed for two experimental and two control conditions, using new recordings of the items used by Norris et al. (2003). Experimental words and nonwords as well as filler words and nonwords were produced by the talker of the pretest and recorded during the same session. Experimental items were 20 [f]-final Dutch words (e.g., *olijf*, 'olive'), 20 [s]-final Dutch words (e.g., *radijs*, 'radish'), and 20 strings that would be nonwords whether they ended in [f] or [s] (e.g., *kwirtaf*, *kwirtas*). Note that *olijs* and *radijf* are not words in Dutch. These three sets were matched in triplets for stress pattern, final vowel, and length (such that there were five items per set with one, two, three, and four syllables). The two real word sets were also matched for frequency (13 per million for [f]-final words and 14 per million for [s]-final words). Except for the final [f] and [s] in the real word sets, no experimental item contained any further instances of these two sounds, nor of [v] or [z]. In addition, there were 80 filler words and 100 filler nonwords, with each of these sets consisting of an equal proportion of items with one, two, three, and four syllables. None of the fillers contained the sounds [f], [s], [v], or [z]. The full set of experimental materials is listed in Norris et al..

There were two versions of each experimental word. One was a natural pronunciation, but in the second version the final fricative was replaced by the ambiguous sound [?] (e.g., *olij?*). To ensure that any transitional information in the final vowel did not cue [f] or [s] and was consistent across sets, ambiguous versions were made from recordings in which the final phoneme was intentionally mispronounced as the velar fricative [x] (e.g., [olɛɪf] as [olɛɪx]). This velar fricative was then excised from the preceding vowel at a zero crossing at the onset of frication, and replaced by [?]. Experimental nonwords were also created from recordings with a final velar fricative.

Phonetic categorization For one pair of experimental exposure groups and the two control exposure groups, the items of the categorization phase were those that had been selected on the basis of the pretest, in the context of a vowel from the same talker (Talker 1). The other pair of experimental exposure groups listened to test stimuli which had been constructed by splicing these same five [f]–[s] steps onto a vowel that had been produced by a different female talker (Talker 2). This vowel [ɛ] was, as with all other spliced items in the experiment, taken from a velar context. A number of tokens of [ɛx] were produced by a female native speaker of Dutch of similar age to Talker 1. Recordings were made in a sound-damped booth onto DAT, then digitally transferred to a computer (48 kHz sampling rate and 16-bit quantization), downsampled to 16 kHz, and edited using Xwaves. A token of [ɛ] (171 ms in duration) was isolated at a zero crossing at the onset of frication and equated in intensity to the vowel of Talker 1 (67.5 dB SPL) before being spliced onto the five fricative steps.

Design and Procedure

Lexical decision There were four exposure conditions, each with 100 words and 100 nonwords. In one experimental condition there were the 20 natural [f]-final words and the ambiguous versions of the 20 [s]-final words (e.g., *olijf* and *radij?*). In addition, there were 60 filler words (15 of each of the four lengths) and 100 nonwords (25 of each length). The second experimental condition was identical except that this list contained the natural versions of the 20 [s]-final words and the ambiguous versions of the 20 [f]-final words (e.g., *olij?* and *radijs*). Two control conditions consisted of the natural recordings of the experimental words, 20 [f]-final items in one and 20 [s]-final items in the other. Listeners in both control conditions also heard the 20 [?]-final experimental nonwords, plus 80 filler words (20 of each length), and 80 filler nonwords (20 of each length).

The 96 participants were assigned to one of six groups (16 participants per group). The two experimental exposure conditions each had two groups which differed only in the stimuli used at test — one that would hear Talker 1 in the test phase and one that would hear Talker 2. The two control exposure conditions each had one group of listeners. Stimuli were presented in a pseudo-randomized running order in which experimental items did not occur on the first 12 trials but were otherwise spread equally across the course of the experiment, with at least four fillers between two experimental items. Running orders for the four conditions were identical to the extent that the appropriate experimental items always appeared in the same positions (i.e., the slot in which one experimental condition contained the natural version of a word would be filled by the ambiguous version of that word in the other experimental condition, and vice versa). The control conditions were based on the experimental conditions such that the natural versions of experimental words were in the same positions, and ambiguous versions were replaced by nonwords. To maintain an equal number of words and nonwords, twenty filler nonwords were replaced with filler words in the control conditions.

Up to four participants were tested at a time in a quiet room and were presented with stimuli binaurally at a comfortable listening level over closed headphones, with an inter-onset interval of 2.6 s. Instructions (given on a computer screen) were to decide as fast and as accurately as possible whether each item was a real Dutch word or not, and to respond by pressing one of two buttons labelled *Ja* ('yes') and *Nee* ('no'). Participants were further told that there would be a short second part for which they would be given instructions on-screen after the lexical decision task. They were therefore unaware, during the lexical decision phase, that they would be tested later on fricative perception. Half of the participants in each condition gave 'yes' and the other half 'no' responses with their dominant hand.

Phonetic categorization The phonetic categorization task followed immediately after the exposure phase and was exactly the same for the six conditions, except that two of the four experimental groups listened to slightly different stimuli, that is, those that were made with a vowel from Talker 2. Six repetitions of each of the five steps from the [ɛf]–[ɛs] continuum were presented at an inter-onset interval of 2.6 s. Order of presentation was pseudo-randomized to ensure that the five steps were spread evenly across the list and no step would occur twice in a row. Participants were given on-screen instructions to press a button labelled 'F' when they heard an [f]-like sound or a button labelled 'S' for an [s]-like sound. Again, the position of the labels was counterbalanced for handedness. Unlike in the pretest there was no practice block.

Questionnaire The participants who listened to Talker 1 in the categorization phase were given a short questionnaire at the end of the experiment in which they were asked open questions as to whether they noticed anything unusual in the lexical decision part of the experiment, and if so, whether they were conscious of taking this into consideration when making their responses in either part of the experiment. Participants who listened to Talker 2 were given two different questions aimed at getting a measure of whether listeners noticed the talker change. The first question asked if any difference between the two parts had been noticed. Unless the spontaneous answer was that there had been a talker change, the second question then asked explicitly if listeners thought that the voices in the two parts were the same or different.

2.2.2 Results

Lexical Decision

Performance in the lexical decision task was used as a criterion for exclusion of participants in the experimental conditions. If participants failed to label at least 50% of experimental words (ambiguous or natural versions) as existing words they were excluded from further analyses (as in Norris et al., 2003, we excluded these participants since, first, given their unwillingness to label the experimental items as words, it is difficult to interpret their categorization data, and second, failure to label unambiguous items as words most of the time indicates poor compliance with the instructions). In the experimental groups that heard ambiguous [s]-final words there were four participants who were below this cut-off point (two in the same-talker and two in the different-talker condition). In one of the groups that heard ambiguous [f]-final words there was one participant below the cut-off (same-talker condition).

The lexical decision data were analysed in order to have a measure of how acceptable the ambiguous items were compared to the natural items, and secondly, how similar (in terms of acceptability) the [?]-final [f]- and [s]-words were to each other. Mixed 2 × 2 analyses of variance (ANOVAs) were performed by subjects and by items on the reaction times (RTs, adjusted to measure from word offset) for 'yes' responses to experimental words and (separately) on the mean percentages of 'no' responses.

The factor Exposure Group (the two experimental conditions) was a between-subjects factor for the subjects analyses and within-subjects for the items analyses while the second factor Final Fricative (whether the original word ended in [f] or [s]) was between-subjects for the items analyses but within-subjects for the subjects analyses. Tests were performed separately for the same- and different-talker training groups. A summary of mean RTs for 'yes' responses to experimental items is given in Table 2.1. Overall, listeners were faster to label the natural versions as words than the ambiguous versions (mean RTs of 188 ms and 240 ms, respectively), where RTs were slowest for the ambiguous [s]-final items. This difference was reflected in the analysis as a significant interaction between the factors Final Fricative and Exposure Group in both the same-talker exposure groups ($F1(1{,}27) = 6.10$, $p < .05$; $F2(1{,}38) = 20.15$, $p < .001$) and the different-talker groups ($F1(1{,}28) = 18.80$, $p < .001$; $F2(1{,}38) = 21.60$, $p < .001$). Neither of the main effects were significant. These results are similar to those obtained by Norris et al. (2003).

Table 2.1 also shows percentages of 'no' responses to experimental items. Listeners were more likely to accept the natural versions as existing words than the ambiguous versions: On average, they rejected 5% of the natural items and 14% of the ambiguous ones. The relatively high percentage of 23% 'no' responses to ambiguous versions of [s]-final words appears to be due mainly to mono- and bisyllabic items. There were four items which 40% or more of all participants responded 'no' to, all of which were mono- or bisyllabic [?]-final [s]-words. Again, this pattern of results replicates Norris et al. (2003).

The overall difference of 9% in 'no' responses to natural vs. ambiguous items was significant in the same-talker groups ($F1(1{,}27) = 29.28$, $p < .001$; $F2(1{,}38) = 8.38$, $p < .01$) and in the different-talker groups ($F1(1{,}28) =$

Table 2.1: Mean Reaction Times and Mean Percentage 'No' Responses in Lexical Decision in Experiments 1–4.

	Experiment 1*		Experiment 2		Experiment 3		Experiment 4	
	RT	%"No"	RT	%"No"	RT	%"No"	RT	%"No"
Natural Fricatives								
[f]-final words	188	2	222	3	173	3	295	1
[s]-final words	187	7	226	3	183	9	250	12
Ambiguous Fricatives								
[f]-final words	209	4	224	2	196	4	280	5
[s]-final words	272	23	265	22	207	20	288	28

Note. Mean reaction times (RTs, in ms, from word offset) are for 'yes' responses only. *In Experiment 1, the data presented here are the combined results across the four experimental groups.

25.93, $p < .001$; $F2(1,38) = 15.28$, g$p < .001$). The main effects were significant in both the same-talker groups (Final Fricative: $F1(1,27) = 58.21$, $p < .001$; $F2(1,38) = 7.73$, $p < .01$; Exposure Group: $F1(1,27) = 14.18$, $p < .005$; $F2(1,38) = 4.26$, $p < .05$) and the different-talker groups (Final Fricative: $F1(1,28) = 50.21$, $p < .001$; $F2(1,38) = 13.76$, $p < .005$; Exposure Group: $F1(1,28) = 7.43$, $p < .05$; $F2(1,38) = 7.66$, $p < .01$). In short, the results from the same- and different-talker groups were very similar to each other and to the results obtained by Norris et al. (2003). Listeners labelled most of the [?]-final items as words.

Phonetic Categorization

The primary data, however, are those from the test phase. The mean percentages of [f] responses to the five continuum steps are plotted for the six groups in Figure 2.2. In the same-talker conditions, participants who heard the ambiguous [?] in [f]-final words during exposure labelled the continuum mostly as [f], while listeners who heard [?] in [s]-final words during exposure categorized most sounds as [s]. Averaged across steps, this constitutes a 41% difference between groups. Listeners in the two control exposure conditions gave intermediate responses. In an ANOVA on the percentage of [f] responses with Step as a within-subjects factor and Training Condition (experimental vs. control) and Fricative Type (natural [f]-final vs. natural [s]-final words at exposure) as between-subjects factors, there was a significant effect of Step (F(4,228) = 28.61, $p < .01$), indicating that the percentage of [f] responses varied overall across the continuum. The three-way interaction of Step, Training Condition, and Fricative Type was also significant (F(4,228) = 3.04, $p < .05$). There was also a significant interaction of Training Condition and Fricative Type (F(1,57) = 9.03, $p < .01$). No other effects were significant.

One-way repeated measures ANOVAs on the same-talker data were performed for direct comparisons of the two experimental conditions, each of the experimental conditions with their respective control condition, and the two control conditions. Crucially, there was a significant difference between the responses of those who listened to natural [f]-final words and ambiguous [s]-final words, and the responses of those who listened to natural [s]-final words and ambiguous [f]-final words (F(1,27) = 13.81, $p < .01$). The comparison between the training condition that listened to natural [f]-final words and ambiguous [s]-final words and its control group (natural [f]-final words and [?]-final nonwords) was significant (F(1,28) = 4.97, $p < .05$), while the difference between the training condition that listened to natural [s]-final words and ambiguous [f]-final words and its control group (natural [s]-final words and [?]-final nonwords) was not significant (F(1,29) = 4.06, $p < .1$). Importantly, there was no significant difference between the two control groups.

For the different-talker groups, categorization of the continuum steps shifted globally toward [s]. Orthogonal to this shift, there was a mean difference of 22% between the two groups. Data from these two groups were analysed together with the two same-talker groups who had received the same exposure conditions. A repeated measures ANOVA on the percentage of [f] responses with Step as a within-subjects factor and Fricative Type (whether listeners heard natural [f] or [s] words at exposure) and Talker Change (whether there was a talker change in the test phase or not) as between-subjects factors was carried out. There was a significant effect of Step (F(4,220) = 27.80, $p <$ .001) and significant main effects of Fricative Type (F(1,55) = 20.10, $p <$ .01) and Talker Change (F(1,55) = 25.67, $p < .01$). Crucially, there was no interaction between these two factors (F(1,55) = 1.86, $p > .05$), suggesting that the size of the difference between the training groups did not differ as a function of whether there was a talker change or not. This was confirmed in a planned comparison of the two different-talker groups, which showed a significant difference ($F(1{,}28) = 6.26$, $p < .05$).

Questionnaire

Same-talker groups In the experimental condition with natural [f]-final and ambiguous [s]-final words, nine participants (64%) reported that they had heard unusual [s] sounds in some items. Typical comments were that the words had not been articulated properly, or that the talker spoke 'with a lisp'. To the question of how this influenced their responses in the lexical decision task, the most common reply was that they were sometimes in doubt about whether to label these items as words or not (four out of the nine). Two participants replied that they became more alert and listened more carefully when they noticed the unusual sounds. The remaining three did not think that their lexical decision responses were influenced by the unusual fricatives. None of the nine participants reported being influenced by the unusual exposure sounds in their categorization responses.

Only one participant from the other three conditions remarked on unusual fricatives, namely that there were 'English th-sounds' in some of the items. This participant was in the control condition that listened to natural [f]-final words and [?]-final nonwords, and did not report being influenced by the presence of these sounds in either of the two parts of the experiment.

Different-talker groups Three participants replied spontaneously that there were different voices in the exposure and test phases. Of the remaining 27, 18 replied 'different' when asked explicitly whether the voice was the same or different in the two parts, seven replied 'same', and two replied 'don't know'. Overall then, 70% of the participants who were included in the final analysis said that there was a different talker, either spontaneously or when asked explicitly.

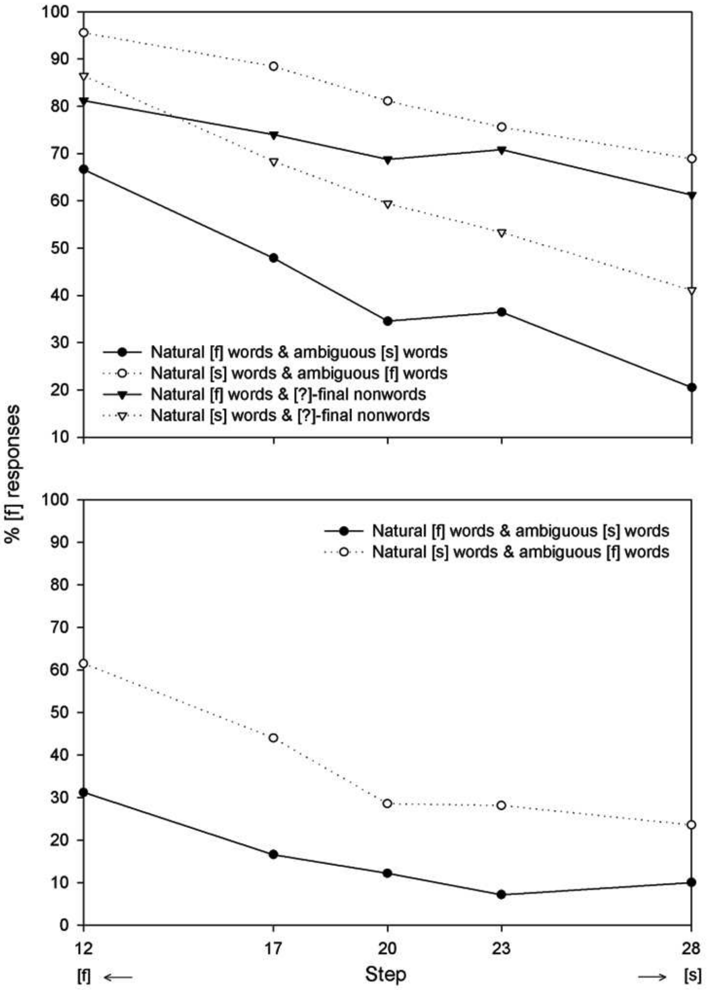

Figure 2.2: Experiment 1: Mean percentage of [f] responses of the six conditions plotted against each of the five continuum steps. Upper panel: Experimental and control exposure conditions with Talker 1's speech presented during exposure and categorization; Lower panel: Experimental exposure conditions with Talker 1's speech presented during exposure, and Talker 2's vowel with Talker 1's fricatives during categorization.

2.2.3 Discussion

The perceptual learning effect reported by Norris et al. (2003) was replicated and found to persist when ambiguous fricatives, made from natural productions by the exposure talker, were presented to listeners in the context of a vowel from a novel talker.

In the exposure phase, the overall performance of the two pairs of experimental groups was very similar. Listeners labelled ambiguous versions of the experimental items as existing Dutch words most of the time. However, although the ambiguous fricative [?] was categorized equally often as [f] or [s] by the pretest listeners, participants in the main experiment seemed to treat this sound more often as an [f]. This [f]-bias was also observed by Norris et al. (2003) and was reflected in a higher percentage of 'no' responses to [s]-final items. The reason for this asymmetry may be that the constant — and therefore uninformative — vocalic context in the pretest encouraged listeners to ignore any coarticulatory cues in the vowel; whereas in the exposure phase the ambiguous fricatives occurred in variable vocalic contexts which apparently cued [f] more reliably than [s] (see Norris et al. for a more detailed discussion).

The main finding of the categorization phase with the same-talker items was a replication of the perceptual learning effect reported by Norris et al. (2003). Listeners who had heard the ambiguous sound [?] in [s]-biased lexical contexts categorized the fricative continuum mostly as [s], while the group which had heard this sound in [f]-biased contexts categorized the same continuum largely as [f]. The control groups, which had been exposed to the same distribution of critical phonemes devoid of lexical context, gave intermediate responses and, as in the Norris et al. study, did not differ from each other. This suggests that, in accordance with Hervais-Adelman et al. (2002) findings on noise-vocoded speech, the observed effect arises as a consequence of lexical feedback and can not be explained by a phonetic contrast effect (i.e., listeners do not appear to be able to learn, on the basis of contrast alone, that since they hear, e.g., an unambiguous [f] during the exposure phase, the ambiguous sound must be an [s]). This lexical influence is related to the Ganong effect (Ganong, 1980) — the tendency of listeners to label ambiguous sounds (including word-final fricatives, McQueen, 1991) in a lexically consistent way. As discussed extensively by Norris et al., however, the present lexical effect differs from the Ganong effect in one crucial respect: It reflects a lexical influence on perceptual learning, rather than a direct influence on explicit phonemic decision-making.

There were two main findings from the conditions in which, during the categorization phase, listeners heard syllables in which the vowel came from a different talker. First, categorization of the continuum shifted towards the [s] endpoint for both groups. This global effect is most likely a consequence of the acoustic properties of the vowel (Johnson, 1991; V. A. Mann & Repp, 1980; V. Mann & Soli, 1991). For instance, one explanation of this shift is that the lower pitch in the vowel (197 Hz for Talker 2 compared to 242 Hz for Talker 1) and/or the lower spectral center of gravity of the vowel (651 Hz for Talker 2 compared to 738 Hz for Talker 1) led listeners to expect a concentration of energy for [f] to occur in a lower frequency region. Since most of the fricatives had energy peaks that were, with respect to the preceding vowel, relatively high in frequency, these sounds were categorized largely as [s]. Borrowing from the literature on vowel normalization, listeners could be said to use *extrinsic* (Johnson, 1990; Nearey, 1989) information to adjust interpretation of linguistic cues in the fricative.

Second, and more importantly, there was again a perceptual learning effect: Listeners who had heard the ambiguous fricative in [s]-biased contexts gave more [s] responses than the other group. This lexically-biased learning effect was orthogonal to the global [s]-bias. There are at least three interpretations of the learning effect. One obvious possibility is that this kind of perceptual learning generalizes to another talker. Listeners in the exposure phase made an adjustment to the [f]–[s] category boundary and this adjustment affected processing of subsequently encountered speech regardless of talker. An alternative explanation is that the effect persists in the different-talker conditions because the fricatives that were used here were still produced by the talker of the exposure phase. It is plausible that these stimuli were recognized by the perceptual system as being produced by the exposure talker and consequently treated as such, even though the preceding vowel indicated that the syllables were produced by a different talker. On this account, the perceptual system analyses the incoming signal for talker identity, and applies previously stored information about the talker on a phoneme-by-phoneme basis. A third account is that using a vowel from a talker of the same sex and similar age did not contain enough information for the perceptual system to treat the utterance as coming from a new talker. For example, Nusbaum and Morin (1992, Experiment 4) found evidence that the speech perception system does not necessarily carry out a new adjustment computation for a new talker if the voice of that talker is acoustically similar enough to that of the previous talker. Although here the majority of participants (70%) indicated hearing a talker change, it is not clear whether this change was processed as such online. Furthermore, only 11% spontaneously pointed out a talker change when they were questioned. When the remaining listeners were asked the question explicitly only very few were confident in their replies. This account was tested in Experiment 2: If the persistent difference in categorization responses between exposure groups observed here is due to too small an acoustic difference between Talker 1 and Talker 2, using a more extreme contrast should eliminate the effect.

2.3 Experiment 2

The aim of this experiment was to test whether the perceptual learning effect that was found for the different-talker groups in Experiment 1 was due to insufficient contrast between the voices of the exposure and the test talker. We thus repeated the different-talker conditions of Experiment 1, but this time the test items were presented in the context of a vowel from a *male* talker.

2.3.1 Method

Participants

Thirty-two volunteers from the MPI for Psycholinguistics participant pool were assigned to two training conditions. None had taken part in Experiment 1, and none reported any hearing disorders. All were paid for their participation.

Materials, Stimulus Construction, and Procedure

Lexical decision Materials in the exposure phase were those used for the experimental groups in Experiment 1.

Phonetic categorization A new set of materials was made for the categorization task in the same way as for the different-talker items in Experiment 1, but this time from recordings of a male native speaker of Dutch (Talker 3). Recording and digitization procedures were as for Talker 2 in Experiment 1. The vowel selected for splicing onto the five fricative steps was 152 ms in duration and equated in intensity to the vowels in the categorization phase of Experiment 1. As before, this [ɛ] was excised from a token of the syllable [ɛx], that is, from a velar fricative context.

Questionnaire Participants were given the same questionnaire as the different-talker exposure groups in Experiment 1.

Procedure The procedures were identical to Experiment 1.

2.3.2 Results

Lexical Decision

We used the same criterion for exclusion of participants as in the previous experiment, which meant that the data from two participants from the group that listened to ambiguous [s]-final words and from one participant from the group that listened to ambiguous [f]-final words were not analysed. The lexical decision data generally show the same pattern as in the previous experiment (see Table 2.1), albeit with some variability across participants' reaction times. Seven mono- or bisyllabic ambiguous [s]-final words were labelled as nonwords by more than 40% of listeners. In the RT data, no effects were significant. In the percentages of 'no' responses, there were significant main effects of both Final Fricative ($F1(1{,}29) = 37.05$, $p < .001$; $F2(1{,}38) = 8.45$, $p < .01$) and Exposure Group ($F1(1{,}29) = 24.29$, $p < .001$; $F2(1{,}38) = 15.55$, $p < .001$), and a significant interaction of the two factors ($F1(1{,}29) = 30.49$, $p < .001$; $F2(1{,}38) = 14.05$, $p < .005$).

Phonetic Categorization

Mean percentages of [f]-responses are given in Figure 2.3. The average difference in responses between the two exposure groups was 25%, in the same direction as in Experiment 1. These data were analysed again together with the two same-talker experimental exposure groups in Experiment 1. There was a significant effect of Step ($F(4{,}224) = 32.17$, $p < .001$), and an interaction of Step and Fricative Type ($F(4{,}224) = 2.99$, $p < .05$). There was also a significant main effect of Fricative Type ($F(1{,}56) = 20.03$, $p < .001$), but, importantly, no interaction of Fricative Type and Talker Change: The difference between the two training groups did not differ as a function of whether there was a talker change or not. Furthermore, a pairwise comparison of the two training groups from Experiment 2 showed a significant difference ($F(1{,}29) = 6.45$, $p < .05$).

Questionnaire

None of the participants spontaneously replied that there was a different voice in the two parts of the experiment; when asked directly however if the talker in the two parts was the same or different, listeners unanimously replied 'different'.

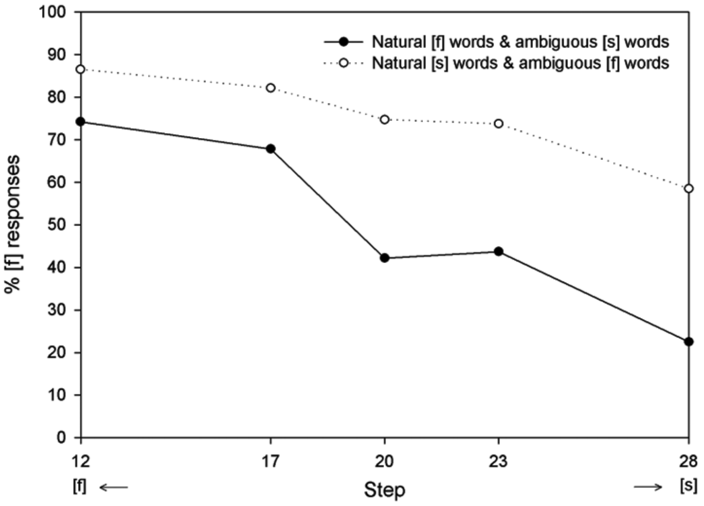

Figure 2.3: Experiment 2: Mean percentage of [f] responses of the two exposure groups to each of the five continuum steps: Talker 1's speech presented during exposure, and Talker 3's vowel with Talker 1's fricatives during categorization.

2.3.3 Discussion

As in Experiment 1, listeners appear to apply a previously learned category boundary shift to fricatives that are presented in the context of a vowel from a novel talker. Unlike in Experiment 1, however, there was no main effect of a change in talker, that is, there was no global shift in categorization responses as a result of a context vowel with different acoustic properties from the vowel of the exposure talker. We suggested earlier that this effect occurred in Experiment 1 because the lower centroid and $F0$ of Talker 2's vowel might have caused a bias to expect the [f]–[s] boundary to be in a lower frequency range too, and consequently led listeners to categorize the continuum largely as (high frequency) [s]. Following this argument, however, a similar shift would be expected in the case of the male vowel produced by Talker 3 in the present experiment since it was even lower in spectral center of gravity and $F0$ (620 Hz and 111 Hz, respectively) than the female vowel produced by Talker 2. Conceivably, there was a more powerful gender normalization process at work which overrode an effect of the kind observed in Experiment 1.

Importantly, we again found an effect of previous exposure even though in this experiment the ambiguous fricatives were presented in the context of a vowel from a male talker, which was acoustically clearly different to the exposure-talker's vowels. Accordingly, the percept that all of our participants reported was that of a male talker during the test phase. Hence, the interpretation of the results of Experiment 1 in terms of insufficient difference between the two talkers used at exposure and test can be dismissed, and we are left with two possible accounts of the present results — namely that the perceptual learning examined here is applied to different talkers, or, alternatively, that the perceptual system 'recognized' the fricative sounds in the test phase as coming from the talker of the exposure phase in spite of the different-talker vowel context, and consequently applied the previously acquired modulation of the [f]–[s] category boundary.

2.4 Experiment 3

In Experiment 3, these two accounts were tested. We used the same experimental exposure conditions as in the previous experiments, but presented listeners with test stimuli in which both the vowel and the ambiguous fricatives came from an unfamiliar talker. If the first account is correct and learning generalizes, we would expect a difference in the categorization responses of the two exposure groups. If, however, learning is talker specific, there should be no effect of exposure in the categorization of fricative sounds from the novel talker.

2.4.1 Method

Participants

Fifty-eight members of the MPI for Psycholinguistics participant pool, none of whom had participated in Experiments 1 or 2, were tested. None of them reported hearing disorders and all were paid for their participation. Forty-eight took part in the main experiment and 10 in a pretest. More participants were tested in the main part than in the previous experiments in order to increase statistical power. Power analysis (Cohen, 1988) after 16 participants in each group had been tested suggested that power was lower by an order of magnitude compared to the same-talker groups in Experiment 1 (0.086 here and 0.842 in Experiment 1), due both to decreased effect size and to increased inter-participant variability in Experiment 3.

Pretest

A pretest was conducted in order to establish five steps on a new [f]–[s] continuum that match the acoustical properties and ambiguity of the stimuli used in Experiments 1 and 2.

Stimulus construction An [ɛf]–[ɛs] continuum based entirely on Talker 3's speech was created using the technique described in Experiment 1. The [f] and [s] endpoints were recorded by Talker 3 in the same recording session as the vowel [ɛ] (which was the token also used in Experiment 2) and re-digitized in the same way. The fricative steps on this new 21-step continuum were matched in duration and intensity to the continuum used in Experiments 1 and 2.

Procedure Informal listening suggested that the most ambiguous range of the continuum was steps 9–18. These ten steps were thus presented to listeners using the same procedure as in the pretest of Experiment 1.

Results Percentages of [f] responses were again averaged for each step. Five steps for the main experiment were selected to match the fricatives used in the previous experiments as closely as possible. Since for the fricatives in those experiments the average percentages of [f] responses were 85, 70, 50, 30, and 15 percent, the steps that corresponded most closely to these percentages were also selected here. To this end it was again necessary to create a more fine-grained 41-step continuum. The five steps that were used for the main experiment, then, were 24, 26, 28, 30, and 32.

Materials and Procedure

The stimuli for the exposure phase were those that were used in the two previous experiments. In the categorization part, the five [ɛf]–[ɛs] steps that were established in the pretest were used. The procedure for the lexical decision and the categorization tasks was as in Experiments 1 and 2, except that participants did not fill in a questionnaire.

2.4.2 Results

Lexical Decision

Application of the 50% cut-off point on lexical decision performance led to exclusion of two participants from the group that listened to natural [f]-final and ambiguous [s]-final words at exposure, leaving 22 participants in that group and 24 in the other. Overall the lexical decision data followed the same pattern as in the previous two experiments (see Table 2.1), again with some variability in the reaction times. There were five ambiguous [s]-final words and one natural [s]-final word that were labelled as nonwords by more than 40% of listeners. In the RT data, there was a significant interaction of Final Fricative and Exposure Group ($F1(1{,}44) = 6.17$, $p < 0.05$; $F2(1{,}38) = 14.28$, $p < .005$). None of the main effects were significant. In the percentages of ‘no’ responses, there were significant main effects of both Final Fricative ($F1(1{,}44) = 61.07$, $p < .001$; $F2(1{,}38) = 6.84$, $p < .05$) and Exposure Group ($F1(1{,}44) = 10.59$, $p < .005$; $F2(1{,}38) = 5.36$, $p < .05$), and a significant interaction between the two factors ($F1(1{,}44) = 20.89$, $p < .001$; $F2(1{,}38) = 9.67$, $p < .005$).

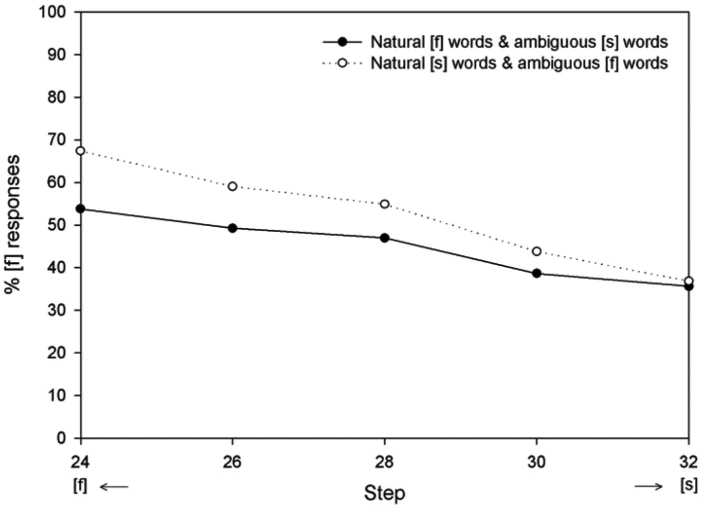

Figure 2.4: Experiment 3: Mean percentage of [f] responses of the two exposure groups to each of the five continuum steps: Talker 1's speech presented during exposure, and Talker 3's vowel and fricatives during categorization.

Phonetic Categorization

The mean percentages of [f] responses to the five fricative sounds are plotted in Figure 2.4. There is a small mean difference of 7% between the exposure groups going in the same direction as in previous experiments. We again conducted a 2 × 2 ANOVA to compare this effect to the categorization data of the same-talker experimental exposure conditions in Experiment 1.

There were significant effects of Step (F(4,284) = 30.88, $p < .001$) and Fricative Type (F(1,71) = 8.79, $p < .005$) but not of Talker Change. Crucially, there was a significant interaction of these two latter factors (F(1,71) = 4.17, $p < .05$), that is, there was a difference in the magnitude of the perceptual learning effect between Experiments 1 and 3. We then conducted a planned comparison of the two exposure groups of Experiment 3 only. There was no significant effect of Fricative Type ($p = .49$), that is, there was a null effect of exposure on categorization responses in this experiment.

2.4.3 Discussion

Unlike in Experiments 1 and 2, where listeners were presented with ambiguous fricatives produced by the talker they had heard in the exposure phase, we here found no effect of exposure when listeners were tested on fricative sounds produced by a novel talker. This result suggests that adjustments to atypical speech are re-applied in a talker-specific manner and do not generalize to processing of utterances from other talkers. Furthermore, the presence of the effect in Experiments 1 and 2 suggests that adjustments affect a specific phonetic contrast, and they are re-applied regardless of the context in which the test sounds appear. Since this conclusion is based on a null effect in Experiment 3, however, it was followed up in Experiment 4.

2.5 Experiment 4

The aim of Experiment 4 was to show that perceptual learning, under appropriate exposure conditions, can be applied to the fricative continuum of Talker 3 that was used in the previous experiment. At the same time we wanted to have another test of the specificity of perceptual learning of a particular phonetic contrast. Would perceptual learning about Talker 3's fricatives occur in the context of words produced by Talker 1?

We therefore used speech editing to splice an ambiguous fricative [?], based on Talker 3's speech, and unambiguous tokens of [f] and [s], into the critical fricative-final materials from the exposure phase, as spoken by Talker 1. Thus, in one version of the materials, the [?] in the [f]-final materials (e.g., *olij?*) spoken by Talker 1 was replaced with the sound used as the most ambiguous step in the Experiment 3 test continuum, based on Talker 3's speech, and the [s] in the [s]-final words (e.g., *radijs*) was a natural [s] spoken by Talker 3. In the other version of the materials the [f] in the [f]-final words came from Talker 3, and the ambiguous sound in the [s]-final words was based on his speech.

If an adjustment of the [f]–[s] boundary on Talker 3's fricatives can be induced by this situation, the null effect in Experiment 3 can be attributed to talker specificity. Furthermore, if learning about Talker 3's fricatives can be induced in the context of speech produced by another talker, this would provide additional support for the account that the perceptual learning mechanism operates regardless of context and can affect a specific phonetic contrast.

2.5.1 Method

Participants

Thirty-nine members of the MPI for Psycholinguistics subjects pool took part. None reported hearing disorders, none had participated in the previous experiments, and all were paid to participate. There were 19 participants in the group receiving [f]-biased exposure and 20 in the group receiving [s]-biased exposure.

Materials and Procedure

Lexical decision There were again two lexically-biased exposure conditions which were identical to those of Experiment 3 in all respects, except for the two following manipulations: The critical ambiguous fricative [?] used for the creation of ambiguous [f]- and [s]-final words was now taken from the Talker 3 continuum (specifically, the fricative sound that had been established as the most ambiguous sound in the pretest of Experiment 3; step 28). Unlike in previous experiments, the natural [f]- and [s]-final words were spliced as well. For these items, the final fricatives were excised at zero-crossings and replaced by the appropriate natural endpoint of Talker 3's continuum (step 1 for [f] and step 41 for [s]).

Phonetic categorization Procedure and stimuli for the test phase were identical to Experiment 3.

2.5.2 Results

Lexical Decision

On the basis of the exclusion criterion used in previous experiments, four participants from the group which listened to natural [f]-final and ambiguous [s]-final words at exposure did not enter further data analyses. The lexical decision data show that subjects tended to respond more slowly than in the previous experiments, and to label more experimental items as nonwords (see Table 2.1). Five ambiguous [s]-final words and two natural [s]-final words were labelled as nonwords by more than 40% of listeners. On average, however, 93% of the natural versions and 84% of the ambiguous versions were accepted as words despite the fact that they had been constructed by concatenating speech from different talkers.

In the percentages of 'no' responses, there was a significant main effect of Final Fricative ($F1(1{,}33) = 64.56$, $p < .001$; $F2(1{,}38) = 6.33$, $p < .05$): [s]-final items were labelled as nonwords more often than [f]-final items. There was no significant main effect of Exposure Group and no significant interaction. In the RT data, neither of the main effects, nor the interaction, were significant.

Phonetic Categorization

The results of the test phase showed that there was a mean difference of 39% in the percentages of [f] responses between the two exposure groups (see Figure 2.5). As in Experiments 1 and 2, the pattern was such that the group which had listened to [?] in [f]-final words gave more [f] responses to the test stimuli.

We first compared this bias effect to the effect in the same-talker conditions of Experiment 1 in a 2 (types of natural fricative at exposure) × 2 (Experiments) ANOVA. There was a significant effect of Step ($F(4{,}240) = 30.92$, $p < .001$) and of Fricative Type ($F(1{,}60) = 29.07$, $p < .001$) but no main effect of Experiment. Importantly, there was no interaction between

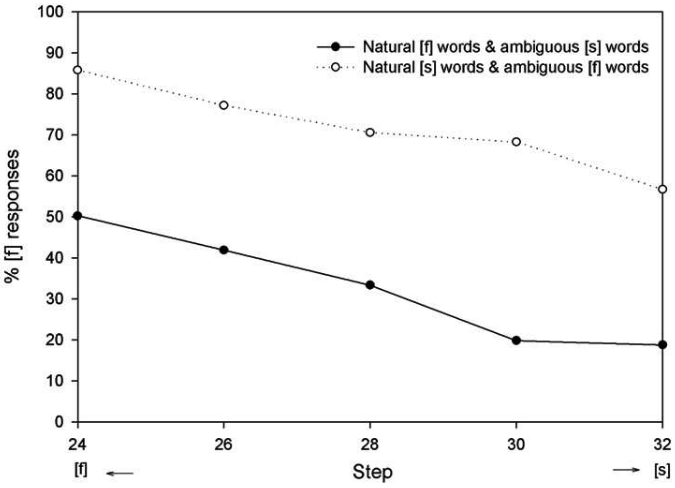

Figure 2.5: Experiment 4: Mean percentage of [f] responses of the two exposure groups to each of the five continuum steps: Talker 3's fricatives in Talker 1's speech presented during exposure, and Talker 3's vowel and fricatives during categorization.

these two factors: The bias effects in the present experiment and in the same-talker conditions of Experiment 1 are of similar magnitude.

Secondly, we repeated this ANOVA with the categorization data from Experiment 3. Again, the only significant main effects were Step (F(4,308) = 36.51, $p < .001$) and Fricative Type (F(1,77) = 9.30, $p < .005$). Crucially, however, the interaction between these factors was significant (F(1,77) = 4.25, $p < .05$). This interaction was then followed up with a pairwise comparison of only the two exposure groups in Experiment 4: F(1,33) = 15.38, $p < .001$.

2.5.3 Discussion

Listeners in this experiment applied an adjustment to Talker 3's fricatives which was learned when an ambiguous fricative produced by Talker 3 was placed in the context of words produced by Talker 1. The learning effect here was statistically indistinguishable from the one in the same-talker conditions in Experiment 1, but different to Experiment 3, where the fricative sounds at exposure and test came from a different talker. We can therefore conclude that the null effect in Experiment 3 was a consequence of the experimental setup, and that it occurred because the perceptual adjustment investigated here does not generalize across talkers.

2.6 General Discussion

The results of this perceptual learning study show that an adjustment made by the perceptual system in response to unusual productions of speech sounds of one talker is stored and re-applied to speech of the same talker, but does not affect processing of speech from other talkers.

Perceptual learning after exposure to an ambiguous fricative sound [?] was evident when this and other sounds on an [f]–[s] continuum were presented in the context of a vowel [ɛ] produced by the talker about whose speech learning had occurred, as well as in the context of vowels produced by other talkers. When presented with test syllables made with vowels from other talkers, listeners perceived a talker change, but the fricatives were treated in a similar way to when they appeared in syllables made entirely from the speech of the exposure talker. With an [ɛf]–[ɛs] test continuum made entirely from utterances of a novel talker, however, we found no evidence of application of previous learning, unless the fricative sounds learned during the exposure phase had themselves originated from the test talker (i.e., the test talker was in fact not entirely new to the listeners).

The perceptual learning effect clearly seems to be lexically mediated (Norris et al., 2003). Evidence for this conclusion comes from two control conditions, in which listeners received the same distribution of critical sounds as the experimental listeners, but in which, unlike in the experimental conditions, ambiguous sounds did not occur in lexical contexts. Since listeners in these control conditions did not show evidence of a category boundary shift, the difference in categorization responses in the experimental conditions can not simply be a contrast effect. Rather, the modulation of the category boundary appears to be the result of a feedback signal from the lexicon. When an incoming ambiguous sound can be disambiguated by lexical information, feedback from the lexicon to a prelexical level results in an adjustment of the phonetic category boundary which can in turn affect perception of future instances of similar ambiguous sounds.

From the perspective of talker normalization, this effect is in line with previous research on normalization of individual's speech (Nygaard et al., 1994; Nygaard & Pisoni, 1998). Listeners make adjustments to idiosyncratic speech production, and the outcome of these computations appears to be stored for later use (Mullennix et al., 1989; Nusbaum & Morin, 1992). One question that was examined here was whether this kind of learning may affect the processing of, or be misapplied to, the speech of other talkers. Given that in Experiment 3 we found no effect of exposure on categorization of ambiguous syllables produced by a novel talker, the answer to this is negative. However, this may turn out to be true only under single-talker conditions. Bradlow and Bent (2003) have shown that listeners are able to apply the outcome of a perceptual adjustment to a novel talker when there are multiple talkers at exposure who share the same idiosyncrasy (in their case, Chinese-accented English). Further, Lively, Logan, and Pisoni (1993) found that talker variability plays an important role in the acquisition of a new phonetic contrast, rather than modification of an existing one. Taken together, these two studies suggest that talker variability facilitates the development and modification of abstract representations of speech. When there are multiple talkers at exposure, the system may be better able to discern acoustic patterns that talkers have in common from those that are idiosyncratic. In the case of single-talker exposure, however, it is less clear which properties of the input signal are characteristic of a phonetic contrast and which are characteristic of the individual talker's vocal tract shape or articulatory habits. The perceptual system would thus be well-advised not to generalize learning to other voices too readily, because such adjustments do not necessarily have a beneficial effect on the processing of other talkers' speech.

Talker specificity in application of perceptual learning is in accord with the results of Mullennix and Pisoni (1990) and Green et al. (1997), who found, at a phonemic level, evidence for a processing dependency between voice information and linguistic information in which linguistic processing is contingent on voice processing. In the present experiments, we found evidence that such a processing dependency exists in the application of a previously learned category boundary modulation. More specifically, our results suggest that application of learned adjustments to a talker is mandatory when that talker's voice is encountered again (i.e., even when that talker's speech sounds occur in the context of another talker's vowels).

A second question we asked concerned the phonetic specificity of perceptual adjustments. The mechanism by which this learning is applied to the incoming speech signal appears to be remarkably sensitive and robust. Given the null effect in Experiment 3, the effect in the talker-change conditions of Experiments 1 and 2 can only be due to the fact that in these experiments fricatives based on the exposure talker's speech were presented. While the syllables as a whole were perceived as coming from a novel talker, the perceptual mechanism which re-applies stored adjustments appears to operate on a sub-syllabic level and irrespective of context. The speech signal thus appears to be monitored continuously for talker identity and for potential useful information about talkers with a resolution at least at the level of individual segments. Additional evidence for a segmental locus of the learning effect was found in Experiment 4. In this experiment a modulation of the [f]–[s] category boundary was made in response to fricatives which were based on the test talkers' speech but had been spliced into the exposure talkers' utterances – there was simply no other information available about the test talker during exposure apart from his [f]–[s] productions.

The present findings thus suggest that the perceptual learning mechanism investigated here affects representations of fricative sounds at a segmental, prelexical level. Further evidence that these adjustments are prelexical comes from a related cross-modal priming experiment (McQueen, Cutler, & Norris, submitted), which used exposure conditions similar to the present experiments. Listeners in that study showed identity priming effects for ambiguous items as a function of exposure condition (ambiguous items such as [doː?] primed either *doof*, ‘deaf’, or *doos*, ‘box’). An adjustment made at a prelexical stage of processing therefore appears to have biased the interpretation of subsequently heard ambiguous sounds, which in turn affected activation of words that had not been heard at exposure.

No current model of word recognition can accommodate perceptual learning at a segmental level. Models which have units of perception only at the lexical level (Klatt, 1979, 1989) can explain adjustments to individual talkers but not specificity of these adjustments at the level of segments. Other models (e.g. McClelland & Elman, 1986; Norris, 1994; Stevens, 2002) which propose abstract phonetic categories prior to lexical access, on the other hand, do not to date have a mechanism of handling talker- (or any other kind of) variability in the process of mapping the incoming speech signal to these categories. They can, however, be supplemented by models specifically aimed at handling variability in the input (e.g. Johnson, 1997; Kruschke, 1992; Nearey, 1989; Smits, 2001, 2002). This will hopefully lead to models of word recognition with increasingly fine-grained and dynamic input representations.

With respect to the relation of talker identity information and phoneme recognition, our results support a processing model in which linguistic and talker identity information are processed in parallel, and where talker identity information constrains the interpretation of linguistic cues (cf. models of vowel normalization, e.g. Hirahara & Kato, 1992). Talker identity information, in this sense, comprises what is intrinsic in the signal and processed on-line, as well as previously acquired and stored information. According to this view of talker normalization, the perceptual system achieves perceptual constancy by exploiting the sources of variability in the speech signal to impose constraints on the interpretation of that inherently ambiguous signal. We therefore do not endorse talker normalization in its narrow sense as a process in which any indexical information is stripped off the signal prior to access to linguistic units of representation (see, e.g. Pisoni, 1997, for discussion).

The results from these experiments extend previous research which has shown that listeners adjust individual phoneme boundaries in response to unusual speech (Ladefoged, 1989; Norris et al., 2003; Scott & Cutler, 1984), and that listeners make talker-specific adjustments (Mullennix et al., 1989; Nusbaum & Morin, 1992; Nygaard et al., 1994) by showing that perceptual adjustments to speech can be highly specific. These adjustments appear to be specific both with respect to segmental information — the adjustments can be specific to a single phonetic contrast — and with respect to information about talker identity — the adjustments can be about one particular talker. We have argued that these findings can best be explained in a model of speech processing in which fricative information is represented at a prelexical stage. These prelexical representations are then modulated by feedback from the lexicon in a talker-specific manner.

References

Bradlow, A. R., & Bent, T. (2003). *Listener adaptation to foreign-accented speech.* Proceedings of the 15th International Congress of Phonetic Sciences, Aug 3–9, Barcelona, Spain.

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*, 521–533.

Clarke, C. M. (2002). Perceptual adjustment to foreign-accented english with short-term exposure. In *Proceedings of the 7th International Conference on Spoken Language Processing, Sept 16–20, Denver, Colorado* (pp. 253–256).

Clarke, C. M. (2003). *Processing time effects of short-term exposure to foreign-accented english* [Doctoral dissertation].

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed).* Hillsdale, NJ: Lawrence Erlbaum.

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

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern british english sentences. *Journal of the Acoustical Society of America*, *115*, 352–361.

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

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*, 1166–1183.

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

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 152–162.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics*, *59*, 675–692.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 421–433.

Hervais-Adelman, A. G., Johnsrude, I. S., Davis, M. H., & Brent, L. (2002). *Adaptation to noise-vocoded speech in normal listeners: Perceptual learning depends on lexical feedback.* Poster presented at the BSA Short Papers Meeting on Experimental Studies of Hearing and Deafness, Sept 16–17, University of Sheffield.

Hirahara, T., & Kato, H. (1992). The effect of $f0$ on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 89–112). Tokyo: Ohmsha.

Johnson, K. (1990). The role of perceived speaker identity in $f0$ normalization of vowels. *Journal of the Acoustical Society of America*, *88*, 642–654.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perception. *Language & Speech*, *34*, 265–279.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Klatt, D. H. (1986). The problem of variability in speech recognition and in models of speech perception. In J. S. Perkell & D. Klatt (Eds.), *Invariance* *and variability in speech processes* (pp. 301–324). Hillsdale, NJ: Lawrence Erlbaum.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Knösche, T. R., Lattner, S., Maess, B., Schauer, M., & Friederici, A. D. (2002). Early parallel processing of auditory word and voice information. *NeuroImage*, *17*, 1493–1503.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Ladefoged, P. (1989). A note on "Information conveyed by vowels". *Journal of the Acoustical Society of America*, *85*, 2223–2224.

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

Lattner, S. (2002). *Neurophysiologische Untersuchungen zur auditorischen Verarbeitung von Stimminformation* [Neurophysiological investigations into auditory processing of voice information]. Doctoral dissertation, Leipzig, Germany. Max Planck Series in Cognitive Neuroscience Vol 29.

Lively, S., Logan, J., & Pisoni, D. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255.

Mann, V., & Soli, S. D. (1991). Perceptual order and the effect of vocalic context on fricative perception. *Perception & Psychophysics*, *49*, 399–411.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception & Psychophysics*, *28*, 213–228.

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*, 676–684.

Maye, J., Aslin, R., & Tanenhaus, M. (2003). In search of the weckud wetch: Online adaptation to speaker accent. Proceedings of the 16^{th} Annual CUNY Conference on Human Sentence Processing, Mar 27–29, Cambridge, MA.

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

McQueen, J. M., Cutler, A., & Norris, D. (submitted). *The mental lexicon is not episodic: A belated reply to Goldinger (1998).*

Mehler, J., Sebastián, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding compressed sentences: The role of rhythm and meaning. *Annals of the New York Academy of Sciences*, *682*, 272–282.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*, 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*, 365–378.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088–2113.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

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

Nusbaum, H., & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–129). San Diego, CA: Academic Press.

Nusbaum, H., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: Ohmsha.

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

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*, 309–328.

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

Pisoni, D. B. (1997). Some thoughts on 'normalization' in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–30). San Diego, CA: Academic Press.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical perception* (pp. 89–112). Cambridge, UK: Cambridge University Press.

Scott, D. R., & Cutler, A. (1984). Segmental phonology and the perception of syntactic structure. *Journal of Verbal Learning and Verbal Behavior*, *23*, 450–466.

Smits, R. (2001). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 1145–1162.

Smits, R. (2002). Hierarchical categorization of coarticulated phonemes. *Perception & Psychophysics*, *63*, 1109–1139.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

Chapter 3

Stability over time

A version of this chapter is currently under review for publication.

3.1 Introduction

When we listen to speech, we need to adjust our interpretation of speech cues in response to talker-specific differences in articulation (Ladefoged & Broadbent, 1957; Ladefoged, 1989). The variability in the speech signal that is introduced by talker idiosyncrasies continues to be problematic for automatic speech recognizers, but is usually handled with remarkable ease by the human perceptual system. By comparing comprehension of novel and familiar talkers under difficult listening conditions, Nygaard, Sommers, and Pisoni (1994) and Nygaard and Pisoni (1998) have shown that being familiar with a talker's voice can even aid comprehension once an initial adjustment has been made.

There are likely to be various processes engaged in perceptual adjustments made to a talker, driven by different sources of talker variability, and operating at several levels, such as the phonemic, lexical, and prosodic levels. A recent study has shown one specific mechanism, which uses lexical knowledge to resolve ambiguities that arise in the signal at the sublexical level (Norris, McQueen, & Cutler, 2003). Exposure to an ambiguous sound [?] that was midway between [f] and [s], caused a shift of the [f]–[s] category boundary when [?] was placed in contexts that were lexically consistent with its interpretation as either [f] or [s]. Two groups of Dutch listeners heard this ambiguous sound while performing a lexical decision task, either in contexts favouring [f] (e.g., *olij?*, where *olijf* is a word, 'olive', but *olijs* is not), or in contexts favouring [s] (e.g., *radij?*, where *radijs* is a word, 'radish', but *radijf* is not). Listeners in the first group subsequently categorized more sounds on an [f]–[s] continuum as [f] than listeners in the second group.

The studies by Nygaard et al. and Norris et al. suggest that the perceptual system has access to previously acquired information about a talker. The present study asks whether this kind of perceptual learning remains stable over a 12-hour period. This follows up on recent research using the Norris et al. exposure–test paradigm that has shown a solid, and under some conditions even increased, perceptual adjustment effect 25 minutes after learning (Kraljic & Samuel, in press-b). A second question was whether conditions that favour consolidation of learning, such that there is little contact with other talkers, as well as the opportunity for sleep, produce a more robust effect than conditions where participants have normal day-to-day interaction with other talkers, and no sleep. A study in which participants were trained to understand synthetic speech has found that, for this type of learning, there is indeed a performance increase when the testing conditions allow sleep over conditions without sleep (Fenn, Nusbaum, & Margoliash, 2003).

To address these questions, an adapted version of the Norris et al. (2003) paradigm was used for inducing a perceptual adjustment. Listeners were first pre-tested on their categorization of [f]-[s] sounds before having lexically-biased exposure to an ambiguous fricative, in the context of passive listening to a story. They were tested again on [f]-[s] categorization immediately after exposure, and after a 12-hour delay, either over the course of one day, or with an intervening night's sleep.

3.2 Method

3.2.1 Participants

Sixty native Dutch speakers with no self-reported hearing disorders took part in exchange for a cash payment. Twenty-four participated in pretests, and 36 participated in the main experiment.

3.2.2 Materials and Stimulus Construction

Speech recordings were made in a sound-damped booth (Sony ECM-MS957 microphone) in a single session and digitized for further processing (Sony SMB-1 A/D converter; 44.1 kHz sampling rate; 16-bit quantization). A female native Dutch speaker produced 20 tokens each of the syllables [ɛf], [ɛs], and [ɛx] for test stimulus construction, and read out two versions of a story (see below) for construction of the exposure materials.

[ɛf]–[ɛs] Continuum

One token each of [f] and [s] was selected from the recorded syllables and excised at zero-crossings at the onset of frication (original durations: [s] 246 ms, [f] 234 ms; original intensities: [s] 67.7 dB SPL, [f] 61.3 dB SPL). The fricatives were cut to a duration of 231 ms, and equated in root mean-square-intensity (62.4 dB SPL). With these sounds as endpoints, an 81-step continuum was made by combining their waveforms in graded, equally spaced proportions (effectively manipulating the spectrum; see McQueen, 1991), where step 1 corresponded to a clear [f] and step 81 to a clear [s]. The resulting fricatives were spliced onto a vowel excised from one of the [ɛx] syllables (duration 111 ms; intensity 79.2 dB SPL). The velar vocalic context was used for all spliced sounds in the experiment in order to avoid transitional cues for [f] or [s].

The [ɛf]–[ɛs] continuum was pretested with 24 Dutch listeners in order to find a maximally ambiguous sound for the exposure materials, and to select stimuli for the test phases of the main experiment. First, twelve listeners categorized ten sounds from the ambiguous range of the continuum (between steps 17 and 53; presented ten times each, in pseudo-randomized order). Using the same procedure, a further twelve listeners then categorized ten stimuli taken from a narrower ambiguous range as determined by the first group's responses (between steps 30 and 53). From the second group's responses, steps on the continuum corresponding to 90, 70, 50, 30, and 10 percent of [f] responses were identified or determined by interpolation. The resulting steps 25, 34, 43, 52, and 61 were used in the test phases of the main experiment. The most ambiguous sound, step 43 ([?]), was also used to create the materials for the exposure phase.

Story

The text of a Dutch translation of a story (Saint-Exupéry, 1943/2001, chapter 2) was edited such that it contained an equal number of [f] and [s] sounds and neither of the sounds [v] or [z] (see appendix A). After editing there were 644 words in total, containing 78 [f] sounds and 78 [s] sounds. Two versions of the story were recorded. In one version, every instance of [f] was intentionally mispronounced as the voiceless velar fricative [x] (e.g., *alsof* 'as if' → [alsɒx]). In the second version every [s] was pronounced as [x] (e.g., *alsof* → [alxɒf]). The 78 critical velar fricatives in both versions were then excised at zero-crossings and replaced by a version of the ambiguous fricative [?]. Since in natural speech the duration of segments is conditioned by various contextual factors, there were three tokens of [?] (all based on step 43). These were made by modifying the amplitude envelope to create two shorter 60-ms and 100-ms sounds (linearly ramped over a 10 ms window at onset and offset), and a long 160-ms sound (ramped over 10 ms at onset and 40 ms at offset). For any given position, the most natural-sounding token out of these three was used. The final two versions of the story were 4.0 minutes long.

3.2.3 Design and Procedure

All participants were given a pretest in which they categorized the five [ɛf]-[ɛs] steps, followed by an exposure phase where the task was to passively listen to one of the two story versions. Immediately after exposure, there was a first categorization posttest, and after a delay of 12 hours, a second posttest.

For 18 participants, the pretest started at 9 am, and posttest-2 was at 9 pm on the same day ('Day group'). For a further 18 subjects, the first session began at 9 pm, while posttest-2 took place at 9 am the following morning ('Night group'). In each of those groups, there were nine listeners who heard the [f]-biased version of the story during exposure (i.e., [?] replacing [f]), and nine listeners who heard the [s]-biased version.

Pretest, posttest-1, and posttest-2 all consisted of ten randomisations of the same five [ɛf]-[ɛs] steps. Stimuli were presented at an inter-onset interval of 2600 ms. Listeners were instructed to press a button labelled 'F' when hearing an [f]-like sound, and a button labelled 'S' for an [s]-like sound.

Table 3.1: Degrees of freedom, F-ratios, and p-values in the analyses of variance of posttest-1 and posttest-2.

		Posttest-1		Posttest-2	
	df	F	p	F	p
Lexical bias	1,33	1.076	.307	2.400	.131
Test	1,33	2.876	.99	4.701	.037
Step	4,132	308.552	6.2e-066	275.888	4.7e-063
Lexical bias × Test	1,33	12.734	.001	8.463	.006
Lexical bias × Step	4,132	3.482	.010	4.722	.001
Test × Step	4,132	3.988	.004	2.564	.041
Lexical bias × Test × Step	4,132	3.781	.006	3.270	.014

3.3 Results

For every test phase, listeners' responses were converted to a percentage of [f] categorizations per step. Data from one participant (Day group; [f]-biased exposure) were discarded since they were at ceiling (100% [f] responses) for every step. All listeners in the Night groups confirmed having had at least six hours of sleep between the posttests.

3.3.1 Immediate learning effect

An immediate learning effect was tested in a mixed analysis of variance (ANOVA) with Test (pretest or posttest-1) and Step as within-subjects factors and Lexical bias ([f]- or [s]-biased exposure) as a between-subjects factor (see Table 3.1). Although there was variability in the individual pretest baselines of the two Lexical bias groups, listeners in the [f]-biased condition showed an increase in [f]-responses from pretest to posttest-1, while listeners in the [s]-biased group showed a decrease (Figure 3.1). Importantly, this interaction of Test and Lexical bias was significant.

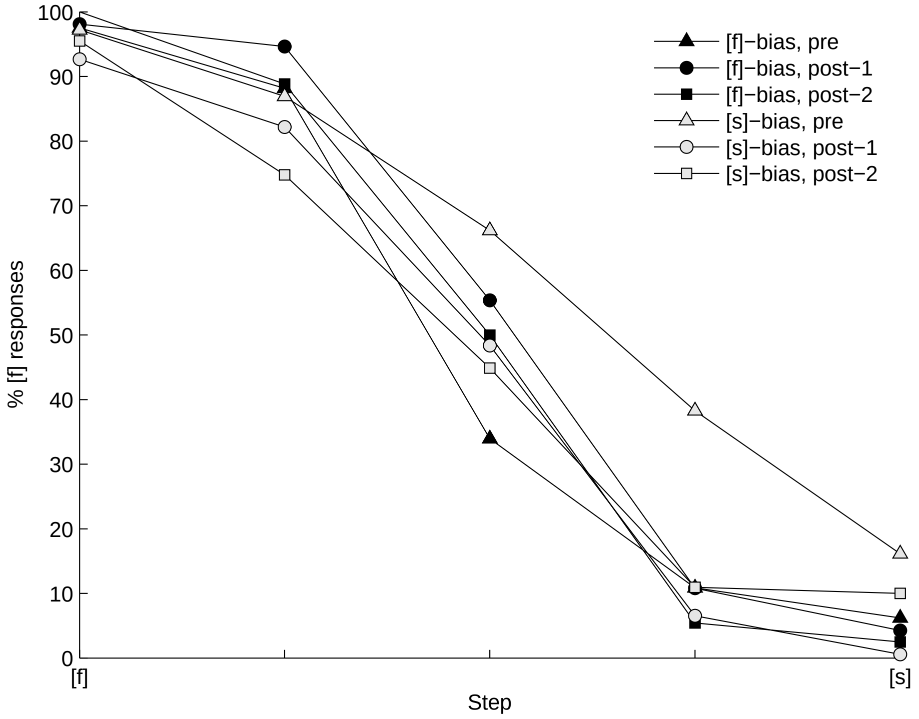

Figure 3.1: Percentages of [f] responses to each of the five [f]–[s] steps for the groups with [f]-biased (filled symbols) and [s]-biased (open symbols) exposure at pretest, posttest-1, and posttest-2.

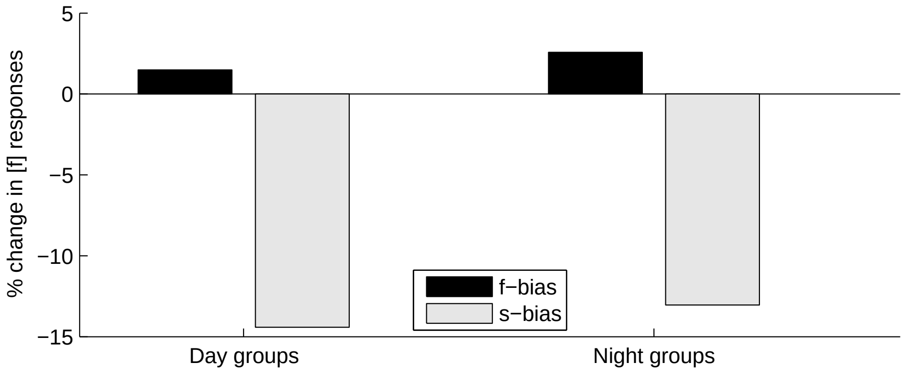

Figure 3.2: Change in percentages of [f] responses from pretest to posttest-2 in the Day and Night conditions (collapsed over Step).

3.3.2 Learning effect after a 12-hour delay

In a similar comparison of posttest-2 and pretest, the interaction of Test and Lexical bias was also significant (see Table 3.1). For a direct comparison of the effects in posttests 1 and 2, we first obtained the root of the squared pretest–posttest differences as an index of effect size (collapsed across Step). These scores were then analysed with Time interval (0 or 12 hours), Lexical bias, and Time of exposure (9 am or 9 pm) as factors. Crucially, there was no main effect of Time interval ($F(1, 31) = .008, p = .930$). No other main effects or interactions were significant (all $p > .25$).

To test for a specific effect of sleep vs. waking on posttest-2 performance (see Figure 3.2), the effect sizes for posttest-2 were analysed separately in an ANOVA with the factors Time of posttest-2 and Lexical bias. Across the exposure groups, there was a small trend towards a greater learning effect for listeners in the Night condition (19% mean shift) compared to those in the Day condition (16% mean shift), but this main effect of Time of posttest-2 was not significant ($F(1, 31) = .775, p = .386$). Again, no other effects were significant (all $p > .25$).[1]

3.4 Discussion

The results show an immediate perceptual learning effect after hearing an ambiguous fricative sound [?] in lexically-biased contexts for a few minutes. In contrast to previous studies using a lexical decision task on a list of words and nonwords as the exposure phase (Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, in press-b), this lexically-guided learning effect was observed here when exposure was passive listening to a short story. Listeners who heard the ambiguous sound placed in words that favour its interpretation as an [f] labelled more sounds on an [f]-[s] continuum as [f] than they did before exposure to [?], while listeners who heard the same sound in [s]-

[1] The learning effect was larger for the groups with [s]-biased exposure. The reason for this asymmetry might be that [f]-like pronunciations of [s] occur outside a laboratory setting more frequently (as a consequence of a speech impediment) than the reverse.

biased contexts showed the reverse pattern. The effect remained robust after a 12-hour interval: No change in magnitude in either direction was observed (relative to the immediate posttest), both for the groups which had the opportunity for consolidation during sleep and received relatively little speech input from other talkers, and the groups which had no sleep and more contact with other talkers.

Fenn et al. (2003) showed that, for learning to understand synthetic speech, there is a decrease in performance during 12 hours of waking but subsequent recovery during sleep. The lack of such a pattern in the present data suggests that the type of perceptual learning examined here is less susceptible to decay. In contrast to learning about synthetic speech, a perceptual adjustment to a talker idiosyncrasy is a very fast-occurring process in which listeners already are highly skilled, and therefore usually unaware of. The perceptual system in this case is not learning a novel skill as such, but applying a subtle adjustment in the processing of a particular phoneme contrast. For this kind of learning to be helpful to the listener in benefiting subsequent recognition of the exposure talker's speech (Norris et al., 2003), it ought to occur rapidly and remain stable regardless of whether the listener is awake or asleep. Although learning to better understand synthetic speech presumably taps into existing prelexical adjustment routines, it is likely to also involve learning at other processing levels (e.g., the unusual prosody of the synthetic 'talker'), all of which may be subject to unlearning during waking. This type of learning also takes time and effort (Greenspan, Nusbaum, & Pisoni, 1988), and often requires explicit feedback during training. It is therefore quite possible that a more drastic distortion of the natural speech signal than the manipulation in the present experiment (e.g., affecting more than one phoneme contrast, or additional levels of processing) will also be more liable to the process of unlearning and recovery that Fenn et al. have demonstrated for synthetic speech.

The picture that is emerging for lexically-driven perceptual adjustments in response to talker idiosyncrasies is that these remain very stable. Using a similar paradigm as the present study, Kraljic and Samuel (in press-b) have already shown that learning effects are reliable after a 25-minute interval, unless listeners are exposed to unambiguous tokens of the critical sound that come from the voice of the exposure talker. Together with these results, the evidence at present suggests that, once the perceptual system has adjusted to a given talker, it does not return to its original state through either the effects of speech input from other talkers or the mere passage of time.

References

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

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*, 614–616.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 421–433.

Kraljic, T., & Samuel, A. G. (in press-b). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology.*

Ladefoged, P. (1989). A note on "Information conveyed by vowels". *Journal of the Acoustical Society of America*, *85*, 2223–2224.

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

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 433–443.

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

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

Saint-Exupéry, A. de. (2001). *De kleine prins* [The little prince] (L. de Beaufort-van Hamel, Trans.). Rotterdam: Ad Donker. (Original work published 1943)

Chapter 4

Where is prelexical processing?

4.1 Introduction

Deriving meaning from spoken utterances requires the evaluation of acoustic cues in the speech signal. These cues are used to access relatively stable, stored representational units of linguistic meaning. The process of getting from the acoustics to word recognition takes place most of the time without any conscious effort by the listener — an indication that the human perceptual system engages highly specialised processes to handle numerous sources of variability in the signal. Many current psycholinguistic models of spoken word recognition include a prelexical level of processing where the speech signal is mapped onto abstract phonetic categories, which in turn pass their activation on to a lexical level of processing (McClelland & Elman, 1986; Norris, 1994; Stevens, 2002; Gaskell & Marslen-Wilson, 1997). Furthermore, models of speech comprehension are often hierarchically organised: Increasingly abstract information flows from early acoustic analysis and prelexical mapping to some kind of perceptual unit (e.g., phonemes, features, diphones) and from there to lexical processing and then higher-order syntactic and semantic levels of processing (McQueen, 2005). The present study addressed the prelexical analysis component of this system. We used functional magnetic resonance imaging (fMRI) to investigate how prelexical processing is implemented in the neuroanatomy of the human brain.

There is evidence for a hierarchical organisation in the neural systems that are involved in the processing of spoken language (Scott & Johnsrude, 2003; Rauschecker, 1998; Kaas & Hackett, 1999; Wise et al., 2001; Davis & Johnsrude, 2003). Cortical processing of a sound starts at the primary auditory cortex (PAC), which occupies the medial two thirds of the transverse temporal gyri, and receives projections primarily from subcortical, ascending auditory pathways. Secondary auditory cortex expands lateral, anterior, and posterior to PAC, and, in humans, may comprise the superior temporal gyrus and the superior temporal sulcus, insular cortex, and the planum temporale (Kaas, Hackett, & Tramo, 1999; Kaas & Hackett, 2000; Rauschecker, 1998; Rauschecker & Tian, 2000). More distant and multimodal regions, including the supramarginal gyrus, inferior frontal gyrus, middle temporal gyrus, and the precentral gyrus, are also frequently seen to be involved in the processing of speech in functional imaging studies. Regions more removed from PAC are often activated in experiments involving higher-level language processing (such as lexical, syntactic, and semantic integration), and include the inferior frontal and inferior temporal gyri as well as the anterior superior temporal sulcus (e.g., Rodd, Davis, & Johnsrude, 2005; Scott, Blank, Rosen, & Wise, 2000; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Davis & Johnsrude, 2003; Sharp, Scott, & Wise, 2003). Functional imaging experiments that employ active tasks (e.g., tasks that require metalinguistic judgements and behavioural responses) often find activation in brain regions that are not typically considered to be receptive language areas (Hickok & Poeppel, 2000; Norris & Wise, 2000; Zatorre, 1997).

While auditory information passes through multiple processing stages in the subcortical auditory pathways and the core regions of PAC (Eggermont, 2001), it is unlikely that any speech-specific processing occurs in these systems. Functional imaging studies have shown that core regions of the auditory cortex are tonotopically organised (Formisano et al., 2003; Yang et al., 2000; Engelien et al., 2002) and respond to pure tones and complex sounds alike. The surrounding cortex, in contrast, is selective for sounds with a more complex spectro-temporal structure. Integration of input from the core areas may therefore take place in these secondary auditory areas (Wessinger et al., 2001). Speech-specific responses in the posterior superior temporal cortex are usually left-lateralised in adults (Binder et al., 1997, 2000; Wise et al., 1991; Scott et al., 2000; Narain et al., 2003) and infants (Peña et al., 2003; Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002), but can also be seen in the right cerebral hemisphere (e.g., after left temporal infarction; Mummery, Ashburner, Scott, & Wise, 1999).

Consistent with this view of a hierarchical organisation of cortical auditory systems, evidence for prelexical processing in brain activation studies is often found in regions that lie lateral to PAC in the superior temporal gyrus and superior temporal sulcus (Scott & Wise, 2004; Indefrey & Cutler, 2004). Magnetoencephalography studies have shown that, in this region, different phonemes elicit discernible patterns in source localisation and latency in the N100m component (Obleser, Elbert, Lahiri, & Eulitz, 2003; Obleser, Lahiri, & Eulitz, 2004). Studies that have attempted to map activations to natural speech sounds with fMRI or positron emission tomography have typically used acoustically-based subtraction designs (e.g., speech vs. Gaussian noise, speech vs. pure tones; Jäncke, Wüstenberg, Scheich, & Heinze, 2002). The conclusions that can be drawn from these types of baseline comparison are limited as they are confounded along other dimensions, such as acoustic complexity, and therefore can often not differentiate between simple acoustic, and speech-specialised processing (Norris & Wise, 2000; Scott & Wise, 2004). This problem has been approached by designing baseline stimuli that are acoustically similar to speech in terms of spectro-temporal complexity, are based on natural speech, and yet are not intelligible utterances (Scott et al., 2000; Narain et al., 2003). Using these types of stimuli in combination with a conjunction design (Price & Friston, 1997), rather than simple baseline subtraction, has identified regions that respond to intelligible speech but not to acoustically complex and unintelligible speech-like sounds in regions on the anterior and posterior superior temporal sulcus.

Other experimental designs address this problem by avoiding a 'static' acoustic baseline subtraction design altogether and instead hold the acoustic signal constant while inducing a change in the phonemic percept. Dehaene-Lambertz et al. (2005) used fMRI to measure cortical activity elicited by sine-wave analogues of spoken syllables — sounds with extremely reduced spectral detail (see Remez, Fellowes, & Rubin, 1997). Their study took advantage of the phenomenon that sine-wave replicas are spectrally so impoverished that they are not perceived as speech by naïve listeners, but can be understood when listeners are told to switch to a 'speech mode'. Perceiving the sine-wave replicas as speech compared to perceiving the same sounds as non-speech produced a left-lateralised activation of the posterior superior temporal gyrus, and the left supramarginal gyrus showed differential activity for different types of speech sounds when listening in 'speech mode'. Using sine-wave analogues of spoken words, Liebenthal, Binder, Piorkowski, and Remez (2003) found a similar, differential fMRI response slightly more ventrally in the left anterolateral transverse temporal gyrus and superior temporal gyrus.

Other studies have attempted to pin down phonological processes against early, nonspecific acoustic analysis by using training paradigms. The rationale in these experiments is that inducing a change along a dimension of interest will show a corresponding change in the fMRI signal relative to a control condition. For example, an initially non-phonemic acoustic pattern, such as an unfamiliar speech sound, can become a perceptual unit after listeners have learned to recognise this sound as belonging to a novel phonemic category. Golestani and Zatorre (2004) trained native English monolingual listeners on a non-native place contrast (retroflex vs. alveolar plosives). They found that only after training did the non-native sounds elicit similar activations to native sounds in areas including both left and right superior temporal gyri, the right middle frontal gyrus and frontal operculum, and the left caudate. Another study addressing acquisition of a non-native phoneme contrast (the [r]/[l] distinction in Japanese listeners; Callan, Tajima, Callan, Kubo, & Akahane-Yamada, 2003), in contrast, found activation of extensive cortical networks to be associated with increased discrimination performance after training. Both the native and non-native contrasts activated superior temporal areas, but the trained nonnative sounds additionally activated frontal, prefrontal, and subcortical areas.

A recent study by Jacquemot, Pallier, LeBihan, Dehaene, and Dupoux (2003) directly investigated native-language phonological processing with fMRI. Instead of a relatively short-term training procedure, this study investigated phonotactics, that is, phonological restrictions that are learned as a result of long-term experience (on the order of years) with a native language. A crossed design was used, with two language groups (French and Japanese) and two phonological contrasts: presence or absence of an epenthetic vowel in a consonant cluster (CVC vs. CC), which is phonologically distinctive in French but illegal in Japanese; and presence of a long vowel (CV:C) vs. a short vowel (CVC), which is a phonological contrast in Japanese but not in French. For Japanese listeners, a CVC sequence is difficult to discriminate from a CC sequence, where they tend to perceive an epenthetic vowel that is not physically there; French listeners, in contrast, find it difficult to distinguish the long and short vowels (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999). Listeners performed an AAX discrimination task in the scanner. The critical comparison was between trials where a 'different' final item constituted a phonological change (i.e., epenthetic vowel for the French, vowel length for the Japanese listeners) and trials where the difference was acoustical (i.e., vowel length for the French, epenthetic vowel for the Japanese; note that the comparison is therefore based on physically identical stimuli). Jacquemot et al. found increased activity for phonological change relative to acoustic change in the left superior temporal and supramarginal gyri, and no activation for the reverse comparison.

In the present study, we used auditory perceptual learning as an approach to identifying brain regions that are engaged in prelexical processing, that is, processing which integrates acoustic cues with attention to contextual factors, and which results in cascaded and probabilistic access of language-specific perceptual representations. As in the Jacquemot et al. (2003) study, the acoustic signal was held constant but the mapping from the acoustics to a more abstract representation was altered by the experimental manipulation. Unlike their study, we examined shifts in the phonetic boundary between two categories rather than a phonotactic effect. Also, unlike previous studies which have employed learning paradigms (Golestani & Zatorre, 2004; Callan et al., 2003) which require relatively intensive and explicit training, we investigated a type of perceptual learning which occurs very fast and without listeners' awareness. We used an adapted version of a paradigm developed by Norris, McQueen, and Cutler (2003), which induces a change in the perception of an ambiguous speech sound. Specifically, in the Norris et al. study, Dutch listeners heard an ambiguous fricative sound that was midway between [f] and [s] embedded in words that favoured the sound's interpretation as either an [f] or an [s] sound (e.g., the sequence *olij?* forms a word in Dutch if the final sound is interpreted as an [f], but not when it is interpreted as an [s]). Listeners who heard this ambiguous sound repeatedly in spoken sequences that are lexically consistent if the sound were an [f], subsequently categorised sounds on an [f]–[s] continuum largely as [f]. A second group of listeners who had been exposed to the same ambiguous sound in contexts that favour its interpretation as an [s] subsequently categorised sounds on the continuum mostly as [s]. Here, we used fMRI to measure brain activity in response to [f]–[s] sounds before and after listeners had lexically-biased exposure. As this type of learning occurs very fast — on the order of a few minutes — pretest, exposure, and posttest all took place within the same scanning session. Behavioural categorisation responses and fMRI images were collected during the pre- and posttest phases; during exposure phase participants listened passively to a story and no images were acquired.

Based on previous research, our first prediction was that regions in primary auditory cortex, as well as the posterior superior temporal gyrus, and potentially the supramarginal gyrus, would be sensitive to the difference in [f]–[s] sounds; with a likely leftward asymmetry for the non-primary areas. Secondly, we predicted that one or more regions identified in this way would show a differential pattern of activation as a function of the lexically-biased exposure, such that a perceptual change would be reflected in evidence for plasticity in the underlying neural systems. Such an effect would allow strong conclusions regarding the localisation of prelexical processing: The initial Norris et al. (2003) study included control conditions which showed that the learning in this paradigm is mainly not acoustic, that is, not due to contrast or selective adaptation effects but driven by language-specific lexical feedback. Furthermore, in another study McQueen, Cutler, and Norris (submitted) have demonstrated that the locus of the adjustment is prelexical, by showing that learning generalised to the processing of lexical items which had not been part of the exposure materials.

Finally, a previous study using this paradigm has shown that the perceptual learning is specific to the voice of the exposure talker (Eisner & McQueen, 2005) and did not generalise when there was a talker change between exposure and test. Kraljic and Samuel (in press-a) have suggested that in addition, the extent of talker-specificity is conditioned by how similar the vocal tract characteristics of the exposure and test talkers are. Given these findings, a secondary prediction for the present experiment was that brain regions which are sensitive to talker-specific information and talker change, such as the left and right middle temporal gyri and the right anterior superior temporal gyrus (Wong, Nusbaum, & Small, 2004; Belin, Fecteau, & Bédard, 2004; Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; Kriegstein & Giraud, 2004) might be activated.

4.2 Method

4.2.1 Participants

Forty-two native speakers of Dutch took part in the experiment. Twenty-four participated in pretests and 18 in the fMRI study. Participants in the fMRI experiment (12 female, 6 male) were right-handed according to the Edinburgh handedness questionnaire and between 19 and 26 (mean 22) years old. None had a history of hearing disorder or neurological illness. All gave informed written consent and were paid for their participation.

4.2.2 Materials and Stimulus Construction

Materials for the test phases were based on an [ɛf]–[ɛs] fricative continuum made from natural speech. Three ambiguous and two relatively unambiguous steps on the continuum were used in the categorisation task. The most ambiguous token of the continuum was used in addition in the exposure materials, where it was inserted in place of [f] or [s] sounds in a continuous speech context.

Speech recordings were made in a single session in a sound-damped booth (Sony ECM-MS957 microphone) and digitized for further processing (Sony SMB-1 A/D converter; 44.1 kHz sampling rate; 16-bit quantization). Two versions of a story (Saint-Exupéry, 1943/2001, chapter 2) were read out by a female native Dutch speaker. These versions of the story formed the basis of the exposure materials. The text of the story had been edited such that it contained an equal number of [f] and [s] sounds (78 of each), and neither of the sounds [v] or [z], embedded in 644 words in total. In one version, the speaker pronounced every instance of [f] as a voiceless velar fricative [x] (e.g., *alsof* 'as if' became [alsɒx]). In the second version every [s] was pronounced as [x] (e.g., *alsof* became [alxɒf]). These were later replaced with an ambiguous [f]-[s] sound; the velar context therefore served to avoid formant transitions in the vowels surrounding the fricatives, which, if they were appropriate for either [f] or [s], could cue the identity of the critical fricatives. In the same recording session, the speaker produced several tokens of the syllables [ɛf], [ɛs], and [ɛx] for the construction of the fricative continuum.

[ɛf]–[ɛs] Continuum

The fricative continuum was made from one token each of [f] and [s], excised from a recorded syllable at zero-crossings at the onset of frication. The original durations were 234 ms and 246 ms for [f] and [s], respectively, and the intensities were 61.3 dB SPL for [f] and 67.7 dB SPL for [s]. The fricatives were cut to a duration of 231 ms, and equated in root-mean-square (RMS) amplitude. These sounds then became the endpoints of an equally-spaced 81-step continuum on which step 1 corresponded to [f] and step 81 to [s], which was made using a linear waveform interpolation procedure (McQueen, 1991). To avoid an intensity confound in the fMRI response (Bilecen, Seifritz, Scheffler, Henning, & Schulte, 2002), all steps were again equated in RMS intensity (62.4 dB SPL) before being spliced onto a vowel which had been excised from one of the recorded [ɛx] syllables (duration 111 ms; intensity 79.2 dB SPL). Again, this velar vocalic context was used to avoid coarticulatory cues for [f] or [s] in the vowel transitions. All speech editing was done with ESPS/Xwaves (Entropic) and Praat (Boersma & Weenink, 2003).

Stimulus selection The [ɛf]–[ɛs] continuum was pretested in order find a maximally ambiguous sound for the exposure materials, and to select stimuli for the test phases.

Participants Twenty-four members of the MPI for Psycholinguistics subject population participated. None reported any hearing disorders, and none took part in the main experiment.

Procedure Listeners were tested individually in a sound-damped booth with the instruction to press a button labelled 'F' when hearing an [f]-like sound and a button labelled 'S' for an [s]-like sound. The first twelve listeners categorised ten randomisations of steps 17, 21, 25, 29, 33, 37, 41, 45, 49, and 53. Responses suggested that the most ambiguous range of the continuum was between steps 30 and 53. A further twelve listeners then categorised stimuli taken from this range (steps 30, 34, 37, 40, and 53; each presented ten times).

Results Responses were converted into a percentage of [f] responses per step. From these data, we determined or interpolated which steps on the continuum corresponded most closely to 90, 70, 50, 30, and 10 percent of [f] responses. The resulting steps 25, 34, 43, 52, and 61 (henceforth [f90], [f70], [f50], [f30], and [f10]) were used in the test phases. As shown in Figure 4.1, the mixing and stimulus selection procedure resulted in a set of test sounds that varied gradually from approximating the spectral shape of a natural [f] to that of a natural [s]. Step 43 ([f50]) was additionally used to build the materials for the exposure phase.

Story

The velar fricatives which had been articulated in place of [f] and [s] in the story recordings were excised at zero-crossings and replaced by the ambiguous fricative [f50]. To account for variability in the duration of segments in natural continuous speech (as caused by multiple factors including phonological context, prosodic context, or speaking rate), there were three versions of [f50] with durations of 60 ms, 100 ms, and 160 ms. These values were based on clusters around these durations in measurements of the natural [f] and [s] sounds in the two story recordings. Duration manipulations were made on

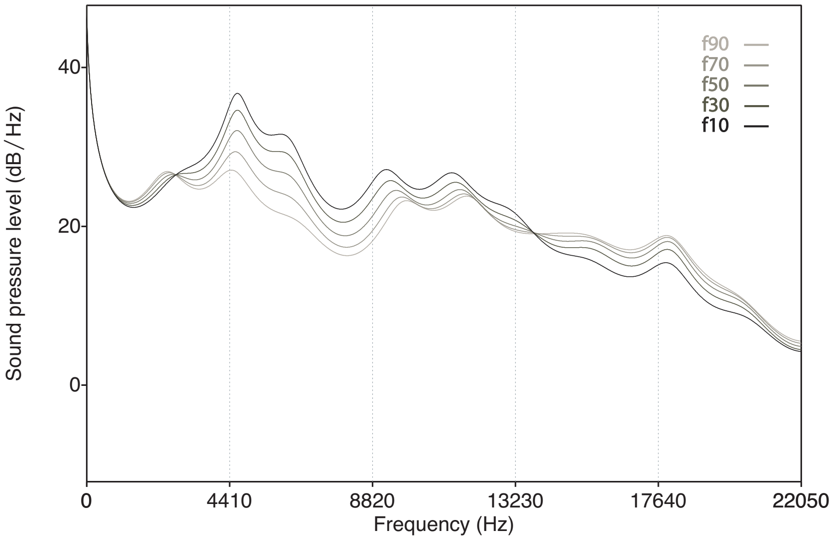

Figure 4.1: LPC-smoothed spectra of the five [f]–[s] test sounds.

the steady-state portion of the fricative. The amplitude envelope was edited, such that the 60-ms and the 100-ms versions were linearly ramped over a 10 ms window at onset and offset, while the 160-ms version was ramped over a 10 ms window at onset and a 40 ms window at offset. For any given position, the token that sounded the most natural was chosen. The final two versions of the story had comparable distributions of the three tokens (ambiguous [f] version: 8 × 60-ms, 66 × 100-ms, 4 × 160-ms; ambiguous [s] version: 13 × 60-ms, 60 × 100-ms, 5 × 160-ms) and were of equal total duration (4 min).

4.2.3 Procedure

Scanning parameters

Functional and anatomical scans were acquired on a 3-tesla system (Siemens TRIO) within the same session. For each subject, two time series of 152 whole-brain images were obtained using a gradient-echo echo-planar imaging (EPI) sequence with prospective acquisition correction for head motion

(Thesen, Heid, Mueller, & Schad, 2000) and the following parameters: 28 axial slices; voxel size 3.5 × 3.5 × 3.5 mm; matrix size 64 × 64 mm; field of view 224 mm; flip angle 90°; echo time (TE) 30 ms; acquisition time (TA) 2 s; and ascending, interleaved slice acquisition. We used a silent event-related paradigm (Amaro et al., 2002; Belin, Zatorre, Hoge, Evans, & Pike, 1999; Moelker & Pattynama, 2003; Hall et al., 1999; Di Salle et al., 2003) with a repetition time (TR) of 10 s to avoid a potential interaction between stimuli and EPI noise (Scarff, Dort, Eggermont, & Goodyear, 2004; Hall et al., 2000). During the silent 8 s interval, one [ɛf]–[ɛs] syllable was presented at one of ten equally spaced stimulus onset times (SOTs), which ranged from 3000 ms to 6150 ms as measured from the offset of the previous scan. Stimulus presentation (controlled by Presentation software; Neurobehavioral Systems) and image acquisition were synchronized with every TR. Stimuli were delivered via earphones (Resonance Technology), which were shielded by circumaural ear defenders and inserted partially into the ear canal. fMRI volumes were collected during each test phase (stimulus presentation began after the second volume), but none during the exposure phase. A structural scan was acquired after the functional runs with a T1-weighted high-resolution sequence (MP-RAGE; 192 sagittal slices).

Categorisation

For the pre- and posttest phases, participants were instructed to press one button when they heard an [f]-like sound and another button for an [s]-like sound. In between the test phases, there was a short exposure phase during which half of the participants passively listened to the [f]-biased version of the story (i.e., [f50] occurred in [f]-positions) and the other half listened to the [s]-biased version (i.e., [f50] occurred in [s]-positions).

Button presses were made with the middle and index fingers of the left hand; button assignments were counterbalanced across participants. The five [ɛf]–[ɛs] steps were presented 30 times each per test phase. The order of presentation was pseudorandomised by concatenating three randomisations of the sounds at each of the ten SOTs (3 × 5 × 10 presentations per test run), with the constraint that no step or SOT occurred more than twice in a row.

Imaging analysis

The MRI data were analysed with BrainVoyager QX (Brain Innovation). The preprocessing steps for the functional images were, in this order, motion correction, slice timing correction, temporal smoothing with a high-pass filter at 4 cycles per second, and spatial smoothing with an isotropic Gaussian kernel of 6 mm full-width-at-half-maximum. The first two functional volumes of each run were discarded, and every participant's functional and structural scans were aligned and transformed into standard stereotaxic space (Talairach & Tournoux, 1988)[1].

Inferential statistics were performed in the context of the general linear model. The model included five predictors of interest corresponding to the five steps of the continuum. The evoked hemodynamic responses were modelled for each event type as stimulus onset (stick function) convolved with a canonical hemodynamic response (i.e., a gamma function, $\delta = 0, \tau = 1.25$; Boynton, Engel, Glover, & Heeger, 1996; Belin et al., 1999). Analyses were performed on the pooled data of all participants where participants were treated as a random factor (Penny & Holmes, 2004). We first conducted an F-contrast, which tests the null hypothesis that *all* parameter estimates are zero. Secondly, t-contrasts were performed in order to identify directly brain regions that are sensitive to the difference between the most [f]-like and most [s]-like sounds. This analysis was restricted to those voxels which were signi-

[1] All reported analyses were also run after aligning the functional and anatomical data with a cortex-based procedure (Fischl, Sereno, Tootell, & Dale, 1999; Goebel, Staedler, Munk, & Muckli, 2002). This type of alignment allows statistical analysis on a reconstructed cortical surface, which is made by first segmenting the cortical sheet in individual subjects from subcortical structures and white matter, and then, through non-linear warping, finding a least-squares solution to match up the individual sheets (encoded as concave and convex curvature values on a spherical space). The procedure aims to increase experimental power by reducing the multiple comparison problem (analyses are run only on cortical voxels) and by improving the inter-subject spatial alignment. The statistical analyses, however, yielded results that were very similar to those obtained in standard stereotaxic space, therefore only the standard-space analyses are reported here.

ficantly activated in the (non-specific) F-contrast (i.e., jointly tested the null hypotheses that all parameter estimates in the F-contrast are zero, *and* that the difference between the estimates for step [f90] and [f10] is zero). In the activated voxels that were identified with this contrast, analyses of variance were then conducted on the regionally pooled beta weights (i.e., the regression coefficients) in order to test for an effect of the lexically-biased exposure conditions in the comparison of pre- and posttest data.

4.3 Results

We discarded the fMRI data from two participants whose behavioural responses were not registered due to a technical error. Two further datasets were excluded from all analyses: those of one participant who failed to respond on more than 50% of trials in the pretest, and of another participant who was unable to distinguish the five test sounds (i.e., showed a flat response function for the continuum). Seven participants remained in each exposure condition.

4.3.1 Behavioural data

Behavioural responses were collapsed into percentages of [f] responses by participant, test phase, and step. The continuum was labelled systematically in both test phases (Figure 4.2). To test for an effect of exposure condition on categorisation performance, the mean difference scores between pre- and posttest percentages of [f] responses were analysed in a repeated measures analysis of variance (ANOVA) with Exposure group ([f]- or [s]-biased) as a between-subjects factor and Step as a within-subjects factor (using a Huynh-Feldt correction for non-sphericity).

Listeners in the group which heard the ambiguous sound in [f]-biased contexts during exposure categorised sounds more often as [f] in the posttest than in the pretest (mean increase of 3% across steps), while listeners in the other exposure condition showed the reverse pattern (mean decrease of 15%).

This main effect of Exposure group was significant (F(1,12) = 18.08, p = .001); no other effects were significant. The effect was less asymmetrical for the most ambiguous step [f50], with a 12% increase for the [f]-biased group and a 28% decrease in the [s]-biased group (univariate ANOVA: F(1,12) = 12.67, p = .004).

A more detailed inspection of the posttest results revealed that the perceptual learning effect was strongest immediately following exposure, and decreased subsequently over the course of the posttest phase. Figure 4.3 shows the mean percentage of [f] responses to step [f50] in the last third of the pretest phase, and the first, second, and final third of the posttest phase. While there was variability in the pretest response levels of the exposure groups, both groups showed a marked shift from pretest-3 to posttest-1, and subsequent decline towards their respective pretest levels in posttest-2 and posttest-3. A statistical analysis showed a significant interaction of Test third and Exposure group (F(5,60) = 3.32, p = .010; both main effects were nonsignificant). In pairwise comparisons of pretest-3 and posttest thirds one, two, and three, only the first two showed (marginally) significant Test third × Exposure group interactions (F(1,12) = 3.79, p = .075; F(1,12) = 9.39, p = .010, respectively); in the final third of the posttest the effect was not reliable any more (F(1,12) = .53, p = .280). No main effects were significant in these pairwise comparisons.

4.3.2 Imaging data

An overall F-test including all effects of interest showed bilateral activation which was strongest in the posterior perisylvian cortex. The three peak clusters in this contrast were on the left and right transverse temporal gyri and, more laterally, on the left superior temporal gyrus. These regions are shown in Figure 4.4 on a statistical parametric map, thresholded using the false discovery rate procedure ($t > 16.0, q(FDR) < .001$; Genovese, Lazar, & Nichols, 2002).

In the t-contrast, four distinct regions showed larger activity for the most [s]-like sound, step [f10], as compared to the most [f]-like sound, step [f90]

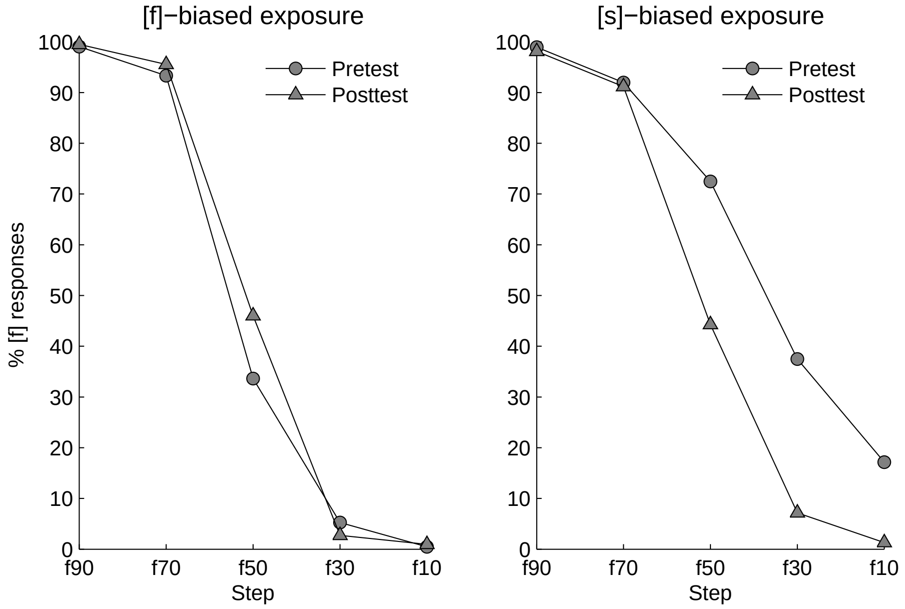

Figure 4.2: Mean percentages of [f] responses across the continuum in the two exposure groups for pre- and posttest.

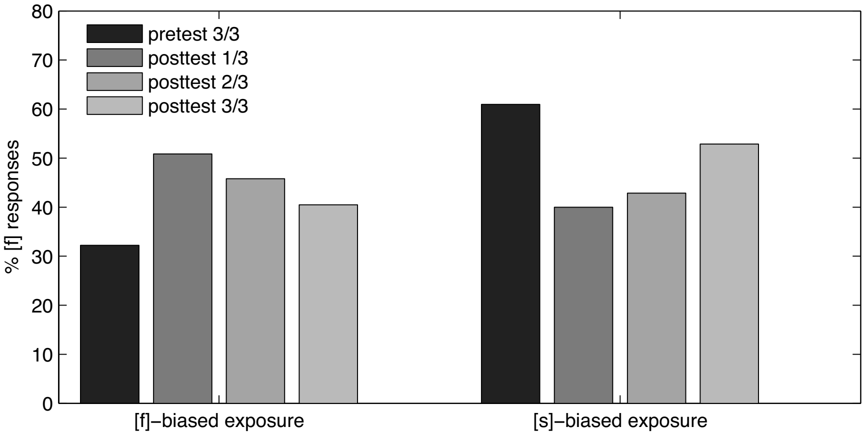

Figure 4.3: Average percentages of [f] responses to step [f50] in the last third of the pretest phase, and thirds one, two, and three of the posttest phase.

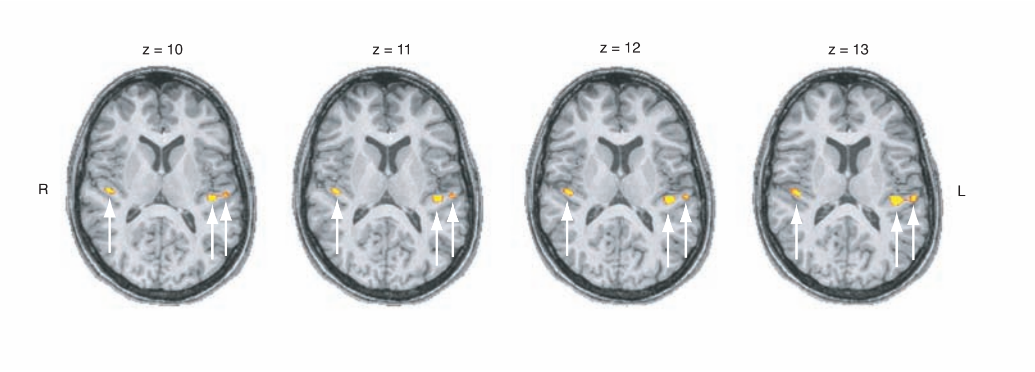

Figure 4.4: Peak activations in the F-contrast ($t > 16.0, q(FDR) < .001$). Locations in the axial plane are indicated in mm as stereotaxic coordinates .

($t > 3.6, p < .005, uncorrected$). No voxels exceeded this threshold in the reverse [step f90 − step f10] contrast. Two of the four activated regions were in the left and right primary auditory cortex (PAC) on the medial transverse temporal gyri. Only on the left was there activation in the superior temporal gyrus lateral to PAC, and only on the right was there activation of the supplementary motor area on the medial superior frontal gyrus. Details of these four regions are given in Table 4.1. Figure 4.5 shows the mean beta weights for the five [f]–[s] steps across participants and test phases.

Since a precise anatomical localisation of function is problematic in auditory cortex, due to a relatively high degree of intersubject variability in this area (Brett, Johnsrude, & Owen, 2002; Rademacher, Bürgel, & Zilles, 2002), we used cytoarchitectonic probability maps in addition to macroanatomical landmarks for determining the location of the four regions of interest. The peak voxel coordinates (transformed into MNI space; Brett, 2002) were compared to probability maps of primary auditory cortex (Eickhoff et al., 2005; Morosan et al., 2001; Rademacher et al., 2001). The results suggested a high probability (80%) for the right temporal peak activation to be in primary auditory cortex. Lying more lateral than on the right, the left medial temporal peak had a probability of only 30% for being located in PAC, and the posterior and lateral peak activation on the superior temporal gyrus had a probability of zero (since currently available maps do not cover the entire brain, this peak could not be assigned to any other region either). The fourth peak on the right superior frontal gyrus was assigned to area 6 (70%) and area 4a (20%).

Finally, we tested whether, as a function of learning, one or more of these regions would show a differential response for the two exposure groups in the comparison of pretest to posttest data. To this end, a repeated measures ANOVA, with Step and Test (pre or post) as within-subjects factors, and Exposure group ([f]- or [s]-biased) as a between-subjects factor, was performed on the pooled beta weights of every subject in a given region (see Table 4.3.2). A learning effect in this analysis would be reflected as a significant interaction of Exposure group and Test, and an additional interaction by Region (e.g., a significant learning effect in STG but not in PAC) would provide the most compelling evidence (Henson, 2005).

Crucially, as Table 4.3.2 shows, the Test × Exposure group interaction was not significant in any of the four regions of interest. There was a significant of Step in all regions but the left transverse temporal gyrus, reflecting sensitivity to the stimulus continuum. Given the un-learning trend that was evident in the behavioural posttest results, this analysis was repeated with only the data from the final third of the pretest and the first third of the posttest. The results were very similar to those for the full dataset, and the Test × Exposure group interaction was again nonsignificant in all four regions.

4.4 Discussion

This study has investigated the neural prelexical processing of speech. We aimed to identify the cortical regions that are implicated in the prelexical mapping of acoustic cues to phonetic categories by using a perceptual learning paradigm. The result of this type of perceptual learning is that identical acoustic cues to an ambiguous speech sound elicit different phonemic percepts, dependent on lexically-biased exposure to this ambiguous sound. Behavioural responses collected during fMRI acquisition showed a reliable effect

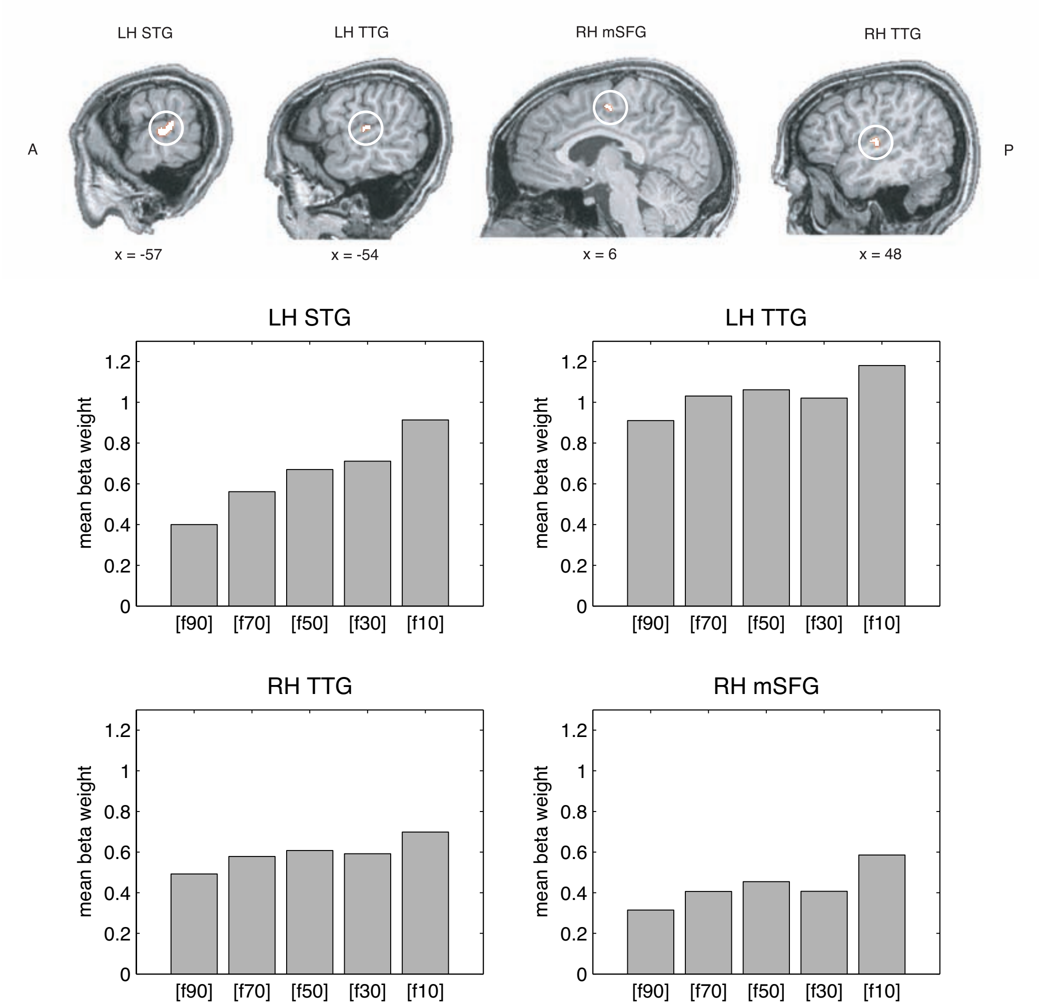

Figure 4.5: *Top.* Regions of interest as identified in the [$stepf10 > stepf90$] t-contrast superimposed on saggital structural images of one participant ($t > 3.6, p < .005, uncorrected$). *Bottom.* Beta weights averaged by region for the five [ɛf]–[ɛs] steps. Data shown are collapsed across test phase and participants. LH, left hemisphere; RH, right hemisphere; STG, superior temporal gyrus; TTG, transverse temporal gyrus; mSFG, medial superior frontal gyrus.

Table 4.1: Location, cluster size in voxels, and stereotaxic coordinates of peak voxels in regions showing larger activity for more [s]-like sounds than for [f]-like sounds .

	BA	No. of voxels	Coordinates		
			x	y	z
Left lateral superior temporal gyrus	22	589	-57	-37	10
Left transverse temporal gyrus	41	71	-54	-25	13
Right transverse temporal gyrus	41	77	48	-13	4
Right medial superior frontal gyrus	6	52	6	-10	49

Table 4.2: Degrees of freedom, F-ratios, and p-values in the analyses of the beta weights in four regions for which activity was higher for [s]- than for [f]-like sounds. LH, left hemisphere; RH, right hemisphere; STG, superior temporal gyrus; TTG, transverse temporal gyrus; mSFG, medial superior frontal gyrus.

		LH STG		LH TTG		RH TTG		RH mSFG	
	df	F	p	F	p	F	p	F	p
Exposure	1,12	.72	.413	.25	.627	.37	.556	.01	.973
Test	1,12	2.02	.181	1.44	.253	1.27	.282	6.12	.029
Step	4,48	9.36	.000	2.00	.109	3.72	.010	6.23	.000
Exposure × Test	1,12	1.53	.240	.10	.754	.25	.625	3.31	.094
Exposure × Step	4,48	.14	.965	.45	.773	1.30	.283	.28	.889
Test × Step	4,48	1.51	.214	2.05	.103	1.62	.186	2.60	.048
Exposure × Test × Step	4,48	1.62	.184	1.66	.175	.06	.992	1.16	.340

of the exposure manipulation. We observed left-lateralised candidate regions that were sensitive to the phoneme contrast of interest, yet none of these regions showed evidence of a learning effect.

In the behavioural results there was a perceptual learning effect (Norris et al., 2003) which by now has been replicated for various phoneme contrasts in a number of studies (Clarke & Luce, 2005; Eisner & McQueen, 2005; Kraljic & Samuel, in press-a, in press-b; Eisner & McQueen, submitted; McQueen et al., submitted; McQueen, Norris, & Cutler, in press). Participants who heard an ambiguous fricative sound embedded in connected speech, and in contexts that lexically favour the sound's interpretation as an [f], categorised more sounds on the [ɛf]–[ɛs] test continuum as [f] in the posttest than in the pretest. Listeners with [s]-biased exposure categorised the same sounds more often as [s] than they did in the pretest. This effect was numerically largest for, but not restricted to, the most ambiguous sound of the continuum. An interesting pattern was the tendency of the perceptual adjustment to partially reverse over the course of the posttest. After a relatively large shift from pretest to posttest, listeners' categorisations receded towards their pretest levels gradually, but not fully. One possible account of this apparent partial reversal of the perceptual learning effect is that the category boundary is re-adjusted upon repeated exposure to relatively unambiguous test stimuli (steps [f10] and [f90]) in combination with a high number of repetitions (note that neither of those design choices would have been warranted in a purely behavioural test setup, but were necessary in order to meet the constraints that fMRI has with respect to contrast sensitivity and experimental power). For instance, listeners who learned during exposure that the test talker produces [f]-sounds in an unusual way, un-learn upon hearing in the posttest that the same talker can actually produce an [f] quite clearly. This effect is consistent with a recent finding by Kraljic and Samuel (in press-b), who conducted a systematic investigation of the processes that reverse this type of perceptual adjustment. They found that only listening to unambiguous tokens of the critical phoneme that were produced by the exposure talker could reverse the initial perceptual learning. It is also unlikely that the partial reversal observed here is simply due to passage of time. In the absence of hearing unambiguous productions from the exposure talker, reliable effects have been reported after intervals of 25 minutes (Kraljic & Samuel, in press-b) and 12 hours (Eisner & McQueen, submitted) after exposure.

The overall F-test of the group fMRI data revealed extensive activation that was strongest in bilateral primary auditory cortex. An interesting result of this analysis was that a separate peak cluster was located lateral to PAC only in the left cerebral hemisphere. Since this contrast tests whether *all* parameter estimates are zero, significant activation in a brain region may be attributed to hearing any of the five speech stimuli. Note that the contrast does not indicate any differences in activation to the different fricative sounds, and that any observed activations could also be driven by the vowel [ɛ] sound in the test syllables, or by the participants performing the categorisation task as such. Although this contrast thus can not pinpoint regions that necessarily distinguish the [ɛf]–[ɛs] test sounds, the left-lateralised non-primary superior temporal cluster may indeed be part of a system that is specialised for the processing of speech. This inference is certainly not conclusive, but the result is consistent with current views of left-lateralised specialised processing streams that extend laterally towards the superior temporal gyrus from PAC (Wessinger et al., 2001; Scott & Johnsrude, 2003; Scott & Wise, 2004; Rauschecker, 1998).

The main interest of the study, however, was in identifying candidate regions in which a learning effect could be observed. T-tests revealed four clusters on the transverse temporal gyri bilaterally, the right superior frontal gyrus, and the left posterior superior temporal gyrus which were differentially sensitive to the endpoints of the range of test sounds. A subsequent analysis of the pre- and posttest beta weights in these four regions confirmed a sensitivity to the stimulus continuum, but none showed evidence of experience-dependent plasticity resulting from the experimental manipulation.

The activation we observed in the superior frontal gyrus lies in the supplementary motor area and is likely implicated in the planning of a behavioural response (i.e., performing the categorisation task), rather than in auditory processing. Although motor areas are sometimes activated during passive listening, these are either seen in premotor cortex (e.g., Wilson, Saygin, Sereno, & Iacoboni, 2004), which is involved in speech production, or in primary motor cortex when the experiment involves semantic processing of 'action' verbs (Pulvermüller, 2005; Hauk, Johnsrude, & Pulvermüller, 2004). Consistent with this interpretation, activity in the superior frontal gyrus was right-lateralised, as can be expected when button presses are made with the left hand.

Current models of the neural architecture of speech processing would suggest that the bilateral PAC activity is likely to be purely a consequence of the spectral differences in the [f]–[s] sounds, and not of their phonemic status. The left-lateralised activation in the posterior superior temporal gyrus, in contrast, is certainly within an area for which there is already evidence for an involvement in prelexical analysis of speech (Hickok & Poeppel, 2000; Scott & Johnsrude, 2003; Scott & Wise, 2004; Davis & Johnsrude, 2003; Jacquemot et al., 2003). This region was also the most extensive, and most significantly activated out of the four, and showed a consistent and gradual response to the stimulus continuum with activity being strongest for the most [s]-like sounds. An additional effect of the lexically-biased exposure would have been very strong evidence that this region is engaged in prelexical processing.

Given this rather positive result, the obvious question is why there was no evidence for a learning effect. A simple explanation may be that the fMRI design did not have enough power to reliably detect the rather subtle change after learning. If indeed the measurements taken here were simply too noisy, one way to improve the sensitivity in a future study might be including an *explicit* baseline condition — perhaps a low-level acoustic manipulation such as signal-correlated noise or spectral rotation. In this way, cortical regions can be revealed that distinguish the test fricative sounds in conjunction with being more sensitive to the speech sounds than to the baseline (i.e., $([f10] > [f90]) \cap (([f10] + [f30] + [f50] + [f70] + [f90]) > \text{baseline})$; and the reverse). On the downside, given that scanning time within one session is limited, there is a trade-off, as an additional condition also implies some loss of power for the experimental conditions of interest. Another possibility for improving sensitivity would be to collect fMRI pretest data from each individual participant in a separate session, which could employ more unambiguous test sounds, or again, an explicit baseline. This might provide enough power to establish *subject-specific* regions that are sensitive to the [f]–[s] contrast, which would then become pre-defined regions-of-interest in a main learning experiment, thereby reducing the problem of macroanatomical inter-subject variability in auditory cortex. A further advantage of this design would be that the main experiment need not include unambiguous sounds and can have fewer trials, both of which, based on the current results, will possibly increase the magnitude of a learning effect.

A more pessimistic interpretation of the absence of a neural learning effect is that it is principally undetectable with fMRI. Possible reasons for this are that the adjustment is implemented in distributed networks, which may not be the same as those that are engaged in prelexical encoding of speech sounds. Furthermore, the neural correlates of perceptual learning in general might involve a variety of neural mechanisms including changes in temporal firing patterns, decreases or increases in the size of receptive cortical fields, and shifting of cortical fields in space (see, e.g., Gilbert, Sigman, & Crist, 2001, for a review). However, other theories on perceptual learning have proposed that learning in neural networks takes place in the very systems that process a given stimulus attribute. Karni and Bertini (1997), for example, argued that "a parsimonious interpretation of the specificity of perceptual learning is that only levels of representation in which a given parameter is differentially represented will undergo learning-dependent changes" (p. 530). Following their notion, perceptual learning in speech is encoded in the brain such that systems engaged in making a phonemic distinction are also engaged when a learned adjustment affects this distinction. In other words, an underlying assumption in the present experiment was that perceptual learning takes place in those neural systems which identify sounds as belonging to contrastive phonetic categories. The present result suggests that fMRI is in principle capable of detecting such a system. Methodological issues along the lines outlined above should be ruled out before the technique itself can be dismissed as too insensitive or unsuitable for investigations into this type of learning.

References

Amaro, E., Jr, R, W. S. C., Shergill, S. S., Fu, C. H. Y., MacSweeney, M., Picchioni, M. M., et al. (2002). Acoustic noise and functional magnetic resonance imaging: Current strategies and future prospects. *Journal of Magnetic Resonance Imaging*, *16*, 497–510.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135.

Belin, P., Zatorre, R. J., Hoge, R., Evans, A. C., & Pike, B. (1999). Event-related fMRI of the auditory cortex. *NeuroImage*, *10*, 417–429.

Bilecen, D., Seifritz, E., Scheffler, K., Henning, J., & Schulte, A. C. (2002). Amplitopicity of the human auditory cortex: An fMRI study. *NeuroImage*, *17*, 710–718.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*, 512–528.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience*, *17*, 353–362.

Boersma, P., & Weenink, D. (2003). Praat 4.1 [Computer software]. Retrieved from [URL] www.fon.hum.uva.nl/praat/.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magentic resonance imaging in human V1. *The Journal of Neuroscience*, *16*(13), 4207–4221.

Brett, M. (2002). *The MNI brain and the Talairach atlas.* [URL] www.mrc-cbu.cam.ac.uk/Imaging/Common/mnispace.shtml.

Brett, M., Johnsrude, I. S., & Owen, A. M. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, *3*, 243–249.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, S., R Masaki, & Akahane-

Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *NeuroImage*, *19*, 113–124.

Clarke, C. M., & Luce, P. A. (2005, July). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. In V. Hazan & P. Iverson (Eds.), *Proceedings of the ISCA workshop on Plasticity in Speech Perception, London, UK, June 15–17* (pp. 23–26).

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience*, *23*(8), 3423–3431.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*, 2013–2015.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24*(1), 21–33.

Di Salle, F., Esposito, F., Scarabino, T., Formisano, E., Marciano, E., Saulino, C., et al. (2003). fMRI of the auditory system: Understanding the neural basis of auditory gestalt. *Magnetic Resonance Imaging*, *21*, 1213–1224.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568–1578.

Eggermont, J. J. (2001). Between sound and perception: Reviewing the search for a neural code. *Hearing Research*, *157*, 1–42.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*, 1325–1335.

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. (submitted). Contraints on perceptual learning in speech: Stability over time.

Engelien, A., Yang, Y., Engelien, W., Zonana, J., Stern, J., & Silbersweig, D. A. (2002). Physiological mapping of human auditory cortices with a silent event-related fMRI technique. *NeuroImage*, *16*, 944–953.

Fischl, B., Sereno, M. I., Tootell, R. B. H., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*, 272–284.

Formisano, E., Kim, D.-S., Di Salle, F., Moortele, P.-F. v. d., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *40*, 859–869.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*, 870–878.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron*, *31*, 681–697.

Goebel, R., Staedler, E., Munk, M. H. J., & Muckli, L. (2002). Cortex-based alignment using functional and structural constraints. *NeuroImage Supplement.*

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage*, *21*, 494–506.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping*, *7*, 213–223.

Hall, D. A., Summerfield, A. Q., Gonçalves, M. S., Foster, J. R., Palmer, A. R., & Bowtell, R. W. (2000). Time course of the auditory BOLD response to scanner noise. *Magnetic Resonance in Medicine*, *43*, 601–606.

Hauk, O., Johnsrude, I. S., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology*, *58A*, 193–233.

Hickok, & Poeppel. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*, 131–138.

Indefrey, P., & Cutler, A. (2004). Prelexical and lexical processing in listening. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (3^{rd} ed., pp. 759–774). Cambridge, MA: MIT Press.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, *23*, 9541–9546.

Jäncke, L., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *NeuroImage*, *15*, 733–746.

Kaas, J. H., & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nature Neuroscience*, *2*(12), 1045–1047.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11793–11799.

Kaas, J. H., Hackett, T. A., & Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, *9*, 164–170.

Karni, A., & Bertini, G. (1997). Learning perceptual skills: Behavioral probes into adult cortical plasticity. *Current Opinion in Neurobiology*, *7*, 530–535.

Kraljic, T., & Samuel, A. G. (in press-a). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review.*

Kraljic, T., & Samuel, A. G. (in press-b). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology.*

Kriegstein, K. von, Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48–55.

Kriegstein, K. von, & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*, 948–955.

Liebenthal, E., Binder, J. R., Piorkowski, R. L., & Remez, R. E. (2003). Short-term reorganization of auditory analysis induced by phonetic experience. *Journal of Cognitive Neuroscience*, *15*(4), 549–558.

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

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 433–443.

McQueen, J. M. (2005). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *The handbook of cognition* (pp. 255–275). London: Sage Publishers.

McQueen, J. M., Cutler, A., & Norris, D. (submitted). *The mental lexicon is not episodic: A belated reply to Goldinger (1998).*

McQueen, J. M., Norris, D., & Cutler, A. (in press). The dynamic nature of speech perception. *Language and Speech.*

Moelker, A., & Pattynama, P. M. T. (2003). Acoustic noise concerns in functional magnetic resonance imaging. *Human Brain Mapping*, *20*, 123–141.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage*, *13*, 684–701.

Mummery, C. J., Ashburner, J., Scott, S. K., & Wise, R. J. S. (1999). Functional neuroimaging of speech perception in six normal and two aphasic subjects. *Journal of the Acoustical Society of America*, *106*, 449–457.

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., et al. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex*, *13*, 1362–1368.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

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

Norris, D., & Wise, R. (2000). The study of prelexical and lexical processes in comprehension: Psycholinguistics and functional neuroimaging. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences.* Cambridge, MA: MIT Press.

Obleser, J., Elbert, T., Lahiri, A., & Eulitz, C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research*, *15*, 207–213.

Obleser, J., Lahiri, A., & Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken words. *Journal of Cognitive Neuroscience*, *16*, 31–39.

Peña, M., Maki, A., Kovačić, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., et al. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11702–11705.

Penny, W. D., & Holmes, A. (2004). Random effects analysis. In K. J. Friston et al. (Eds.), *Human brain function* (2nd ed.). London: Academic Press.

Price, C. J., & Friston, K. (1997). Cognitive conjunction: A new approach to brain activation experiments. *NeuroImage*, *5*, 261–270.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*, 576–582.

Rademacher, J., Bürgel, U., & Zilles, K. (2002). Stereotaxic localization, intersubject variability, and interhemispheric differences of the human auditory thalamocortical system. *NeuroImage*, *17*, 142–160.

Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H. J., et al. (2001). Probabilistic mapping and volume measurement of human primary auditory cortex. *NeuroImage*, *13*, 669–683.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, *8*, 516–521.

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11800–11806.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 651–666.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*, 1261–1269.

Saint-Exupéry, A. de. (2001). *De kleine prins* [The little prince] (L. de Beaufort-van Hamel, Trans.). Rotterdam: Ad Donker. (Original work published 1943)

Scarff, C. J., Dort, J. C., Eggermont, J. J., & Goodyear, B. G. (2004). The effect of MR scanner noise on auditory cortex activity using fMRI. *Human Brain Mapping*, *22*, 341–349.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*, 2400–2406.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*, 100–107.

Scott, S. K., & Wise, R. J. S. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, *92*, 13–45.

Sharp, D. J., Scott, S. K., & Wise, R. J. S. (2003). Monitoring and the controlled processing of meaning: Distinct prefrontal systems. *Cerebral Cortex*, *14*, 1–10.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* Stuttgart: Thieme.

Thesen, S., Heid, O., Mueller, E., & Schad, L. R. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magnetic Resonance in Medicine*, *44*, 457–465.

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *13*, 1–7.

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

Wise, R. J. S., Chollet, F., Hadar, U., Friston, K., Hoffner, E., & Frackowiak, R. J. S. (1991). Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain*, *114*(4), 1803–1817.

Wise, R. J. S., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., & Warburton, E. A. (2001). Separate neural subsystems within 'Wernicke's area'. *Brain*, *124*, 83–95.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*, 1173–1184.

Yang, Y., Engelien, A., Engelien, W., Xu, S., Stern, E., & Silbersweig, D. A. (2000). A silent event-related functional MRI technique for brain activation studies without interference of scanner acoustic noise. *Magnetic Resonance in Medicine*, *43*, 185–190.

Zatorre, R. J. (1997). Cerebral correlates of human auditory processing. In J. Syka (Ed.), *Acoustical signal processing in the central auditory system.* New York: Plenum Press.

Chapter 5

Summary and General Discussion

The last few years have been an interesting time to be working on plasticity in the speech perception system. There have been many new developments in diverse fields — a recent ISCA workshop dedicated solely to this topic had contributions from researchers working in areas including phonetics, psycholinguistics, cognitive neuroscience, infant development, cochlear implantation, second language acquisition, and automatic speech recognition. All of these lines of research are concerned with the changes that occur when the speech perception system encounters input that is in some way novel. The experiments reported in this thesis investigated a type of perceptual learning which allows the adult perceptual system to dynamically adjust to unusual speech productions of a kind that may frequently occur in natural listening situations. The experiments were based on a series of studies by Norris, McQueen, and Cutler (2003), who showed that listeners use stored lexical knowledge to adjust prelexical processing in response to an ambiguity in the acoustic-phonetic signal. In line with Gibson's (1969) definition of perceptual learning, the function of this process is presumably adaptive — it can help the listener to decode more efficiently the message intended by a talker whose productions of a given speech sound are outside of a prototypical range. The primary causes for idiosyncratic realisations may include, for example, unusual vocal tract characteristics, a speech impediment, or an unfamiliar accent. However, lexically-guided learning might also be implicated in other domains, for example when infants acquire a native language, adults learn a foreign language, or hearing-impaired individuals adapt to the spectrally degraded input from a cochlear implant. In this chapter, the main findings of the thesis are discussed in the context of other recent research on perceptual learning in speech.

5.1 Specificity

The experiments in chapter 2 tested whether a modulation of the [f]/[s] category boundary resulting from lexically-biased exposure to an ambiguous fricative is specific to the talker whose ambiguous productions caused the adjustment, or whether there is generalisation when listeners hear speech coming from other talkers. For the adjustment to be useful to the listener, it should only be applied again when speech from the exposure talker is encountered, so that the adjustment does not have to be carried out over and over again. It is less likely to be beneficial when applied indiscriminately to any member of the listener's language community, as long as there is no evidence that others share the idiosyncrasy in their speech production. The results of the experiments suggested that perceptual learning induced in this way is indeed highly talker-specific: Listeners applied the category boundary modulation only to fricative test sounds uttered by the exposure talker. Effects of equal magnitude were observed even when when these sounds were presented in the context of carrier vowels from other male and female talkers which elicited the percept of a talker change. No effect was found with test fricatives that were produced by a novel talker. An effect was observed, however, when, under identical test conditions, this novel talker's ambiguous fricatives had been spliced into the original talker's speech production during exposure.

An issue which is related to this talker-specificity, and which has not been tested with this paradigm yet, is whether exposure to multiple talkers who share the same idiosyncrasy in their productions will be more likely to produce generalisation. There are some recent findings by Bradlow and Bent (2003) which suggest that this might be the case. In their study, English listeners were better able to identify words spoken in Chinese-accented English after they had had exposure to multiple talkers with that accent than with equivalent exposure to only a single talker. In the exposure phase of this study, listeners heard sentence-length utterances; therefore the possibility can not be excluded that the adjustment that resulted in better intelligibility affected rhythmic or prosodic processing rather than the prelexical processing that was examined in the current experiments. However, it is conceivable that a similar lexically-driven adjustment mechanism also operated in Bradlow and Bent's study, but that after two days of relatively intensive exposure, listeners developed a cognitive representation of the Chinese accent which was no longer specific to the exposure talkers. Most of us know from subjective experience that this kind of learning occurs, for example when moving to a new dialectal environment, although the effect becomes noticeable more likely on the order of days and weeks than within minutes. (Evans & Iverson, 2004) have shown that listeners indeed maintain representations of different accents of their language, and that recognising a familiar accent can introduce perceptual biases in the on-line analysis of the speech signal. Again, if the type of learning investigated here turned out to generalise after multiple-talker exposure, this would serve the adaptive purpose of facilitating future word recognition.

Two recent studies which also used a variant of the Norris et al. (2003) exposure–test paradigm have reported data that may further qualify the conditions under which talker-specific learning occurs. Kraljic and Samuel (in press-a) found that the perceptual learning effect generalises to speech input from other talkers when the phonetic category modulation affected a stop contrast, rather than a fricative contrast. They tested native English speakers after lexically-biased exposure to an ambiguous [d]–[t] sound on both [d]–[t] and [b]–[p] categorisation. Listeners categorised two versions of these continua, made on the basis of speech from two different talkers. The continua were created by manipulating primarily the temporal cues to stop identification — voice onset time (VOT) and burst duration — as well as intensity of the burst. Unlike in the experiments in chapter 2, there was therefore no spectral manipulation, neither of the release burst nor of the surrounding vowel context. The observed generalisation to a novel test talker (when the results were collapsed across the different place continua) suggests that a perceptual adjustment to a temporal cue is not used by the perceptual system in a talker-specific manner. This is surprising as it is also known that listeners encode this level of detail (VOT) for individual talkers into memory (Allen & Miller, 2004). So the question then arises why the perceptual system does not appear to use information that could presumably benefit comprehension.

A second interesting finding of Kraljic and Samuel's study was that there was also generalisation to the bilabial stop continua (when collapsing responses across talkers). Generalisation effects between bilabial and alveolar place of articulation for VOT have already been reported in other studies, both for learning a novel VOT contrast (Tremblay, Kraus, Carrell, & McGee, 1997; McClaskey, Pisoni, & Carrell, 1983) and for selective adaptation (Eimas & Corbit, 1973). For the case of these stop consonants, the perceptual adjustment may thus mainly affect a voicing cue which is relatively abstract. More specifically, a speculative interpretation of this generalisation is that the temporal VOT cue is adjusted at a higher level in the perceptual system than the low-level spectral manipulation of the kind that was employed for fricatives in chapter 2 (Kraljic & Samuel, in press-a). However, a further recent study from another group which also used the Norris et al. (2003) paradigm, found no evidence for generalisation from a trained alveolar stop contrast to a velar [g]–[k] contrast (Clarke & Luce, 2005). This latter result raises the possibility that lexically-driven VOT adjustments do not generalise indiscriminately, and may interact with other spectral cues in the speech signal.

A second recent study by Kraljic and Samuel (in press-b) investigated the conditions under which perceptual learning might be reversed; this time using a [s]/[ʃ] fricative contrast. They again employed a variant of the Norris et al. (2003) paradigm, but here with a 25-minute interval in between exposure and test. The time interval by itself produced no decrease of the effect (but in some conditions an increase). Hearing either a talker other than the exposure talker produce unambiguous tokens of the critical trained sounds between exposure and test, or hearing the exposure talker produce speech that contained none of the critical sounds, also had no effect on the magnitude of perceptual learning. Hearing the exposure talker produce unambiguous versions of the trained sounds during the 25-minute interval, however, did significantly reduce the effect. This pattern was obtained both for male and female voices, and is in line with the findings in chapter 2, as it suggests talker-specificity of perceptual learning.

As a second measure of talker-specificity, all of these conditions were also run with a talker change in the test phase. The combined results of the talker change conditions were asymmetrical, such that conditions with a male talker at exposure and a female talker at test showed no perceptual learning effect, suggesting talker-specificity, whereas hearing the female talker during exposure and the male talker in the test phase did show an effect, suggesting generalisation. Kraljic and Samuel proposed that the latter conflicting result was caused by an asymmetry in the average spectral centre of gravity of the exposure and test stimuli, as was revealed in a post-hoc acoustic analysis. First, the exposure and test items for the male voice were more similar to each other (average difference of ∼5 Hz) than for the female voice (average difference of ∼640 Hz). Second, because also the absolute centroid values were different (all higher for the female voice), the case of male exposure/female test represented a larger difference (of ∼1160 Hz) than in the case of female exposure/male test (∼520 Hz difference). The explanation for generalisation of perceptual learning put forward by Kraljic and Samuel is therefore based on acoustic similarity between the exposure and test sound. If correct, the specificity of perceptual learning is not an all-or-nothing phenomenon but of a more gradual nature, such that a perceptual adjustment to unusually produced speech sounds of one talker will also be applied to other sounds that have somewhat similar acoustic characteristics, regardless of who produced the sounds. Because of some differences in design and stimulus construction, and of the different fricative contrast that was used, this account cannot be tested directly against the results of chapter 2. However, an equivalent analysis of our stimuli does not appear to support it: The difference between the male and female fricative sounds in chapter 2, which produced no generalisation, was even smaller (∼320 Hz) than the difference in the Kraljic and Samuel materials for which generalisation was observed. The question of how ‘acoustically similar’ test stimuli would need to be in order to produce or not produce generalisation can only be addressed in a systematic investigation of this account.

In chapter 2 another aspect of the specificity of this type of perceptual adjustment was addressed, namely that of the speech cues affected by learning. Specifically, Experiment 4 tested whether an adjustment is made with respect to the processing of cues extrinsic to the critical phonetic category, such as general vocal tract characteristics, or, alternatively, to cues intrinsic to the category. Critically, in this experiment the ambiguous sound at exposure as well as the test syllables were produced by a talker other than the talker who had uttered the lexical carriers. A perceptual learning effect of equal magnitude to that in previous single-talker versions of the experiment was found. Since there was no information about the test talker present at exposure other than spectral cues of the ambiguous segments, the result suggests that this type of perceptual adjustment operates primarily on category-intrinsic cues. The same conclusion follows from a comparison of two conditions in the Kraljic and Samuel (in press-b) study: when the exposure talker produced unambiguous versions of the critical fricative sounds after learning, the adjustment was undone, whereas when the same talker produced speech which did not contain any of the critical sounds, no effect on the adjustment was observed. Since the speech context in both conditions was closely matched, these results support the notion that lexically-driven perceptual adjustments of the category boundary affect the category itself, and that an involvement of category-extrinsic parameters is likely to be negligible.

The current data on the specificity of lexically-guided perceptual learning in speech suggest that, while there may be situations in which generalisation occurs (e.g., after multiple-talker exposure, or when learning adjusts a more abstract featural representation), there are clear cases which demonstrate that learning can be talker-specific. Talker-specific knowledge has been shown to affect the processing of fine phonetic detail which in turn affects the phonetic category boundary between two speech sounds, and must therefore be stored or accessed in some way by the perceptual system.

5.2 Stability

Chapter 3 investigated whether a lexically-driven adjustment to the phonetic category boundary is a short-lived phenomenon, that may be maintained only for a short duration and then be discarded, or, whether these adjustments remain stable over time. Two groups of listeners were either exposed to manipulated speech in the morning and tested 12 hours later in the evening of the same day, or had the exposure phase in the evening and were then tested on the following morning. All participants were also tested immediately after exposure, which served as a baseline measure. Two factors in this design may in principle produce a more stable effect for the group for which learning took place in the evening. First, these participants were very likely to receive much less speech input from other talkers which would have contained unambiguous productions of the critical speech sounds, and therefore might have corrected the initial adjustment. Second, the participants in this group had slept for at least six hours before they were tested again the next morning. There are some parallels to a previous study by Fenn, Nusbaum, and Margoliash (2003), in which listeners were trained on transcribing poorly synthesised speech. This study found that listeners showed improved performance after training, and this effect decayed over the course of a day but not over the course of a night during which participants had slept. There are, however, also some important differences which make such an influence of sleep on the lexically-driven learning effect less likely. Unlike the adjustment to synthetic speech in the Fenn et al. study, the learning in these experiments took place without any explicit training and generally without listeners' awareness. Accommodating an unusual pronunciation of a speech sound presumably reflects a process which listeners engage very frequently and which is thus, in contrast to the rather unfamiliar synthetic speech sounds, highly overlearned. For this learning to be useful to the listener, it should not require a lot of time to consolidate.

The results showed that indeed there was significant perceptual learning immediately after training, and this effect decreased neither for the groups that had the exposure in the morning, nor for those that had it in the evening. There was also no difference between the two groups in the magnitude of the effect after a 12-hour interval. These results suggest that perceptual learning remained very stable during the tested period, and that there is neither a decay during waking due to interference from other talkers, nor an additional benefit from having the opportunity for consolidation of learning during sleep. The results are in line with the study by Kraljic and Samuel (in press-b) that was described earlier — they too found no effect of a novel talker producing unambiguous tokens from the critical phoneme contrast, and they also found the effect to remain stable during an interval of 25 minutes. While the Fenn et al. (2003) experiments employed stimuli which had undergone an unfamiliar type of distortion, the spectral modification of a fricative sound as used in the present thesis is a manipulation that, while artificial, nonetheless falls within the range of sounds that could be produced by a human vocal tract, and is not unlike the kind of variability that occurs naturally between talkers. It seems plausible that because the perceptual system is highly experienced with this kind of variability, the lexically-driven adjustment of a phonetic category is both very rapid and stable.

5.3 Attention

All experiments in chapter 2, as in the original Norris et al. (2003) study and others that have used this paradigm (McQueen, Cutler, & Norris, submitted; Kraljic & Samuel, in press-a, in press-b), used a lexical decision task for the exposure phase. The dual purpose of this task was to keep listeners engaged, and to obtain a measure of whether items with an embedded ambiguous sound would be acceptable as real words for the participants (which was generally the case except for a few monosyllabic items). The use of the lexical decision task raises the possibility that perceptual learning only occurs when listeners' attention is focused on the stimulus attribute of interest, that is, the learning pertains to a feature that is directly relevant to the experimental task. This is arguably the case in lexical decision in this situation — the interpretation of the ambiguous sound is directly relevant to the task because it decides whether the listeners' response will be 'word' or 'non-word'. In the visual modality, task-relevancy has indeed often been reported to be required for perceptual learning to occur (e.g., Ahissar & Hochstein, 2002). When subjects are trained on the discrimination of, say, a subtle difference in the orientation of a line array where the lines also vary subtly in length, a perceptual learning effect is likely to be observed afterwards for orientation but not for length judgements. If participants are trained with an identical stimulus set on the length distinction, perceptual learning would be expected for the length dimension but not for orientation.

A recent study by McQueen, Norris, and Cutler (in press) addressed this issue for perceptual learning in speech. They used exactly the same exposure items, words and nonwords, as in the two critical lexical bias conditions of the Norris et al. (2003) study. In the exposure phase, these items were presented in blocks; the listeners' task was, instead of doing lexical decision, to simply count the number of items in each block, words and nonwords alike. They were also instructed to press a button after each item. A perceptual learning effect was observed that was statistically indistinguishable from that in the original experiment. An analysis of the button-press reaction times revealed further that responses after the items with an embedded ambiguous sound were slower than for those with only unambiguous sounds. This latter result suggests that the ambiguous sounds were noticed at some stage in the perceptual system although they were not relevant to performing the task at all. More importantly, the perceptual learning effect measured after exposure shows that learning does not depend on a rather artificial laboratory task, but occurs in an automatic fashion in response to mere exposure to an ambiguous sound in an appropriate lexical context. This finding is in line with the results from chapter 3 and chapter 4. In those experiments, the exposure materials were a read-out story in which ambiguous fricatives replaced [f] or [s] sounds in various morphological and prosodic positions. The task for the listeners was to follow the story; their attention therefore was, as in real life, on the content of what was being said rather than the identity of individual sounds (and very few listeners noticed the presence of any unusual sounds at all).

The conclusion that can be drawn on the basis of multiple experiments which did not use lexical decision during exposure is therefore that the perceptual system adjusts to unusual pronunciations automatically and does not require focused attention to a specific stimulus attribute. This may reflect a more general characteristic of the speech perception system — intelligible speech in the environment is very hard to ignore, even if we (try to) have our attention elsewhere, and speech seems to be processed from prelexical up to syntactic and semantic levels automatically. The fact that learning occurs in the absence of attention may not be too surprising given that, unlike for novel and abstract discrimination tasks (e.g., of line orientation or pure tone pitch), the speech perception system already has access to all the required mechanisms. Lexically-driven learning is not novel in the sense that it is based on existing lexical representations and an existing learning mechanism, that is, a training signal which originates from those representations.

Because speech captures attention to content so strongly, it might be nearly impossible to tease apart which aspects of the signal are attended to and which are not during listening to continuous and intelligible speech. Whether this type of learning may even occur during sleep (as has been shown for the acquisition of phonetic categories in infants; Cheour et al., 2002) is an interesting empirical question and possibly the ultimate test for automaticity. However, learning during sleep would not be predicted by a recent model proposed by Seitz and Watanabe (2005), which, for the adult visual system, outlines a mechanism which elegantly integrates task-relevant perceptual learning with task-irrelevant learning. Some studies have reported that indeed perceptual learning can occur for an unattended stimulus property also in the visual system, even if that property is barely detectable (e.g., Seitz & Watanabe, 2003). The model proposes that task-irrelevant learning occurs if the unattended stimulus attribute coincides temporally with an attended one, and that a reinforcement signal generated for the attended feature then also facilitates learning for the unattended feature. If applied to speech perception and the kind of perceptual learning that was investigated in this thesis, this model can thus explain why learning occurs both when attention is focused on the critical stimulus attribute, as in the case of the lexical-decision experiments, and when attention is focused on content, as in the passive listening experiments in chapters 3 and 4.

5.4 Neural systems

Previous research on the functional neuroanatomy of speech perception has identified candidate cortical regions in which prelexical processing of the speech signal may occur. In line with psycholinguistic models that posit hierarchical organisation in spoken word recognition, such regions have been proposed to lie lateral to, and receive projections from, the primary auditory cortex, which is known to engage in relatively nonspecific analysis of auditory information. From those lateral superior temporal regions, information may be passed on to areas that lie more anterior on the superior temporal gyrus and the superior temporal sulcus for lexical and semantic processing (Indefrey & Cutler, 2004; Narain et al., 2003; Scott, Blank, Rosen, & Wise, 2000; Scott & Wise, 2004). The experiment in chapter 4 used functional magnetic resonance imaging and an adapted version of the Norris et al. (2003) perceptual learning paradigm to further investigate how prelexical processing is implemented in the neuroanatomy of the brain. The rationale of the experiment was that precisely the kind of change in prelexical processing that is induced by perceptual learning would provide strong evidence for a functional localisation — if correlates of it could be detected with neuroimaging. The prediction was thus that areas which are sensitive to a given phoneme contrast would show a differential response if this contrast is modulated in one or the other direction by lexically-driven learning. Four regions were identified which responded to the critical [f]/[s] contrast; most notably a region on the left posterior superior temporal gyrus, which was consistent with various other studies that have specifically addressed prelexical processes (e.g., Jacquemot, Pallier, LeBihan, Dehaene, & Dupoux, 2003). However, the experiment failed to find any evidence of a change in the neural response as a function of lexically-driven learning in any of these regions. This negative result raises methodological questions which are discussed in chapter 4. In particular, the choice of a baseline condition and the optimal level of ambiguity of the test stimuli are relevant parameters which may be explored in future research.

5.5 Other types of lexical learning

While all experiments in this thesis are concerned with learning that results in an adjustment of a phonetic contrast, there may be important links to other types of learning in speech. One particularly interesting case from a clinical setting is the much more drastic perceptual adjustment that cochlear implant (CI) users need to make to accommodate spectral degradation and distortion of the entire speech signal. The poverty of the stimulation that CI users receive results both from the way the speech is transduced by the CI's processor, and commonly from an imperfect alignment of frequencies along the CI's electrode array with frequency-selective regions along the cochlea. CI users, and normal-hearing listeners who are trained on simulations of this kind of signal, can learn to overcome the distortion, albeit with varying degrees of success (Oh et al., 2003; Rosen, Faulkner, & Wilkinson, 1999).

One study that has investigated which kind of information drives this learning has found that it is again lexical knowledge which is crucial for listeners to adapt to such a signal (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Davis, Johnsrude, et al. trained normal-hearing listeners on noise-vocoded speech, which simulates one aspect of the poverty of the signal available to CI users, namely its reduced spectral resolution (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Compared to frequency misalignment, low spectral resolution is a relatively minor distortion: in Davis, Johnsrude, et al.'s study, for example, even untrained listeners achieved ∼20% correct word identification from only a few spoken sentences. With some experience, and especially with explicit feedback, performance levels rose by 40 to 80% in less than half an hour. Crucially, the results showed that when different types of information available during a pre-exposure phase were compared with respect to the impact they had on performance in a test phase, only lexical knowledge was found to make a significant contribution. In contrast, the comprehension benefit gained from the phonological, syntactic, and semantic levels of representation in isolation was either weak or absent. Davis, Johnsrude, et al. suggested furthermore that the results were most compatible with a locus of adjustment at the level of prelexical representation, as only roughly half of the words that occurred during pre-exposure were also presented in the test phase, suggesting a generalisation across the lexicon that can best be explained by prelexical adjustments. This has since been tested directly in experiments in which listeners were trained on noise-vocoded spoken words and tested on a set of novel words, which again produced transfer of learning (Hervais-Adelman, Davis, & Carlyon, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2005).

In cases of cochlear implantation where listeners have to accommodate additional spectral shifts, or where other factors cause very low initial performance, the use of stored lexical knowledge may be more limited because prelexical processing of the speech signal may perform too poorly to achieve even weak lexical activation. Especially adult CI recipients often need to rely on lipreading at least during the first weeks and months following implantation, but often reach a level of performance where visual information is not required anymore for comprehension. The recent work by Bertelson, Vroomen, and de Gelder (2003) that was described in Chapter 1 may have identified an explicit mechanism for this process by showing that the use of visual cues can result in a modification of prelexical processing of the acoustic-phonetic signal over time. It seems likely that different learning mechanisms need to contribute in order to overcome extreme distortions of the speech signal such as in the case of cochlear implantation.

The studies on learning to understand noise-vocoded speech therefore raise the possibility that the kind of lexically-driven learning that was first shown by Norris et al. (2003), and which has been investigated in this thesis, may reflect a more general capability of the perceptual system. The perceptual system might be able to utilise mechanisms which exist for dealing with natural variability in order to adapt to a more drastic signal distortion in an unnatural listening situation. The extent to which such a learning mechanism may be implicated in other domains, such as infant language acquisition or second language acquisition — as well differentiating lexically-guided learning from other known learning mechanisms in speech — is a topic of currently ongoing research.

5.6 Models of spoken word recognition

There are two main conclusions from the recent perceptual learning literature which have important implications for models of speech perception, as none of the existing models can account for both of these at the same time. The first conclusion is that there is flow of information from the lexical to the prelexical processing level, which can over time modify the mapping of the acoustic signal to prelexical representations if the speech input in some way falls outside the prototypical categories of the listener (Norris et al., 2003; Davis, Johnsrude, et al., 2005). The second conclusion is that at least for some types of prelexical categories, the modification is specific to the talker who produced the unusual speech (Chapter 2; Kraljic & Samuel, in press-b) — models of speech perception therefore require a mechanism for maintaining or accessing talker identity information. More specifically, the experiments in chapter 2 suggested that the perceptual system monitors the information in the speech signal continuously for fine-grained acoustic features that may be specific to a familiar talker, and applies previously learned adjustments to the evaluation of category-intrinsic cues.

On the one hand, some existing models, such as TRACE (McClelland & Elman, 1986) and the distributed cohort model (DCM; Gaskell & Marslen-Wilson, 1997), might be able to accommodate a lexically-driven adjustment since they allow top-down flow of information (on-line in TRACE; during training in DCM). Neither of them, however, have prelexical representations that are detailed enough to account for learned talker-specific effects. As has been discussed in chapter 1, there may be other and possibly more realistic ways to model the learning mechanism computationally, although these have not yet been implemented (Norris et al., 2003; McQueen, 2003).

Episodic models, on the other hand, can in principle account for talker-specificity since this level of detail is still encoded in lexical representations (e.g., Goldinger, 1998). Furthermore, the question of top-down feedback does not apply since there is no lower level for lexical information to feed back to. Models of the kind proposed by Goldinger or Klatt (1979) have no mechanism for an adjustment at a prelexical level and are for this reason incompatible with the recent literature on perceptual learning. One important reason why hierarchical and abstractionist models include prelexical representations is that a change in one such representation will affect processing in all subsequent stages which receive input from it. Learning, in short, can generalise across the lexicon. That generalisation indeed occurs across the lexicon for a prelexical modification has been shown for the case of learning to understand noise-vocoded speech (Davis, Johnsrude, et al., 2005; Davis, Hervais-Adelman, Taylor, Carlyon, & Johnsrude, 2005; Hervais-Adelman, Davis, & Carlyon, 2005), as well as in a recent experiment by McQueen et al. (submitted) for the case of modulation of a single category boundary of the type investigated in this thesis. McQueen et al. used exposure conditions which were equivalent to the Norris et al. (2003) study but then paired with cross-modal priming in the test phase. The results of this study showed that after exposure, an item such as [doː?], which is lexically consistent in Dutch both with [doːs] ('box') and with [doːf] ('deaf'), primed responses to a visually presented DOOS for listeners who previously had had [s]-biased exposure, whereas responses to DOOF were primed in the group with [f]-biased exposure. None of the words in the priming phase had been part of the exposure phase, which strongly suggests that the perceptual adjustment induced during exposure must have affected a prelexical stage of processing, allowing the adjustment to transfer to other words in the lexicon.

The experiments in this thesis and other recent studies on perceptual learning in speech have described a mechanism which enables the perceptual system to adapt rapidly to changing listening situations and thereby maintain perceptual constancy for the listener. By showing that stored lexical representations can modulate prelexical processing over time, and that this process requires access to an on-line analysis of the identity of the talker, they have identified new constraints for models of spoken word recognition. Mechanisms for influencing prelexical processing — both by sending training signals from the lexicon and by accessing previously acquired talker-specific information — should be reflected in the architecture of future models.

References

Ahissar, M., & Hochstein, S. (2002). The role of attention in learning simple visual tasks. In M. Fahle & T. Poggio (Eds.), *Perceptual learning.* Cambridge, Ma.: MIT Press.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115*(6), 3171–3183.

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

Bradlow, A. R., & Bent, T. (2003). *Listener adaptation to foreign-accented speech.* Proceedings of the 15^{th} International Congress of Phonetic Sciences, Aug 3–9, Barcelona, Spain.

Cheour, M., Martynova, O., Näätänen, R., Erkkola, R., Sillanpää, M., Kero, P., et al. (2002). Speech sounds learned by sleeping newborns. *Nature, 415*, 599–600.

Clarke, C. M., & Luce, P. A. (2005, July). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. In V. Hazan & P. Iverson (Eds.), *Proceedings of the ISCA workshop on Plasticity in Speech Perception, London, UK, June 15–17* (pp. 23–26).

Davis, M. H., Hervais-Adelman, A., Taylor, K., Carlyon, R. P., & Johnsrude, I. S. (2005). *Transfer of perceptual learning of vocoded speech: Evidence for abstract pre-lexical representations.* Poster presented at the ISCA workshop on Plastiticy in Speech Perception, London, UK, June 15–17.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222–241.

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

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern british english sentences. *Journal of the Acoustical Society of America*, *115*, 352–361.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*, 614–616.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* Englewood Cliffs, NJ: Prentice-Hall.

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

Hervais-Adelman, A. G., Davis, M. H., & Carlyon, R. P. (2005). *Perceptual learning of noise vocoded words.* Poster presented at the ISCA workshop on Plastiticy in Speech Perception, London, UK, June 15–17.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., & Carlyon, R. (2005). *Perceptual learning of noise-vocoded words: Evidence for top-down processes in speech perception.* Poster presented at the Meeting of the Experimental Psychology Society and the Canadian Society for Brain, Behaviour and Cognitive Science, Montreal, Canada, July 14–17.

Indefrey, P., & Cutler, A. (2004). Prelexical and lexical processing in listening. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (3^{rd} ed., pp. 759–774). Cambridge, MA: MIT Press.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, *23*, 9541–9546.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Kraljic, T., & Samuel, A. G. (in press-a). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review.*

Kraljic, T., & Samuel, A. G. (in press-b). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology.*

McClaskey, C. L., Pisoni, D. B., & Carrell, T. D. (1983). Transfer of learning of a new linguistic contrast in voicing. *Perception & Psychophysics*, *34*(4), 323–330.

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

McQueen, J. M. (2003). The ghost of Christmas future: Didn't Scrooge learn to be good? Commentary on Magnuson, McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science*, *27*, 795–799.

McQueen, J. M., Cutler, A., & Norris, D. (submitted). *The mental lexicon is not episodic: A belated reply to Goldinger (1998).*

McQueen, J. M., Norris, D., & Cutler, A. (in press). The dynamic nature of speech perception. *Language and Speech.*

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., et al. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex*, *13*, 1362–1368.

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

Oh, S.-H., Kim, C.-S., Kang, E. J., Lee, D. S., Lee, H. J., O Chang, S., et al. (2003). Speech perception after cochlear implantation over a 4-year time period. *Acta Otolaryngologica*, *123*, 148–153.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 106*(6), 3629–3636.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*, 2400–2406.

Scott, S. K., & Wise, R. J. S. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, *92*, 13–45.

Seitz, A., & Watanabe, T. (2003). Is subliminal learning really passive? *Nature*, *422*, 36.

Seitz, A., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences*, *9*(7), 329–334.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech perception with primarily temporal cues. *Science*, *270*, 303-304.

Tremblay, K., Kraus, N., Carrell, T. D., & McGee, T. (1997). Central auditory system plasticity: Generalisation to novel stimuli following listening training. *Journal of the Acoustical Society of America*, *102*(6), 3762–3773.

Appendix A

Exposure materials in chapters 3 and 4

Ik leefde alleen, zonder ooit met iemand echt te
kunnen praten, totdat ik op een keer, vijf jaar geleden,
motorpech kreeg in de Sahara-woestijn. Er was wat kapot
gegaan binnen in mijn motor, en omdat ik geen monteur en
ook geen bemanning aan boord had moest ik proberen om
helemaal alleen een moeilijke reparatie uit te voeren. Het was voor
mij van essentieel belang. Ik had nauwelijks
voor vijf dagen drinkwater bij me.
De eerste nacht sliep ik dan ook in het zand, vele honderden mijlen ver
van de bewoonde wereld af. Ik voelde me meer verlaten dan een
schipbreukeling op een vlot midden op de oceaan.
Je kunt je daarom mijn verwondering wel indenken,
toen ik bij het aanbreken van de dag gewekt werd
door een grappig klein stemmetje.
Het zei: “Toe – teken ‘n schaap voor me.”
– Hé? –
– Teken een schaap voor me –
Ik veerde op, alsof de bliksem mij getroffen had –
Ik deed mijn ogen goed open en keek nog een keer. En ik zag een
héél uitzonderlijk klein kereltje, dat me ernstig aankeek.
Ik bekeek die verschijning met ogen die rond van verwondering
waren. Vergeet niet, dat ik honderden mijlen van de bewoonde
wereld af was. Maar dat kleine ventje zag er niet uit, alsof hij
verdwaald was, of doodmoe of hongerig, of dorstig of angstig.

Hij had niets van een verloren kind in de woestijn, honderden
mijlen van de bewoonde wereld af.
Toen ik eindelijk een woord kon uitbrengen, vroeg ik hem:
"Wat doe je hier eigenlijk?" En toen herhaalde hij héél zacht,
alsof het om wat ernstigs ging "Toe, teken 'n schaap voor me."
Wanneer wij van het geheimzinnige al te gemakkelijk onder de indruk raken,
moeten wij wel doen wat opgedragen wordt. Hoe nutteloos het mij ook leek,
honderden mijlen van de bewoonde wereld af en in doodsgevaar,
haalde ik een blaadje papier en een vulpen uit mijn broekzak. Maar toen
bedacht ik, dat ik vooral geografie, filosofie, rekenen
en taal geleerd had en ik vertelde, een beetje humeurig, aan
het kereltje, dat ik niet tekenen kon. Hij antwoordde:
– Dat doet er niet toe. Teken maar 'n schaap voor me.
En omdat ik nog nooit een schaap getekend had, maakte ik
een van de twee enige tekeningen
die ik kon voor hem. De olifant in de boa constrictor.
En ik hoorde hem zeggen:
– Nee, nee! ik wil geen olifant in 'n boa. 'n boa constrictor
is veel te gevaarlijk en een olifant neemt een heleboel ruimte in.
Ik woon erg klein. Ik heb 'n schaap nodig. Teken nou een schaap voor me".
Toen tekende ik het dan maar.
Hij bekeek het aandachtig en klaagde:
– Nee, dat schaap is nù al erg oud. Maak er nog maar een.
Ik tekende. Mijn vriendje lachte vriendelijk en toegeeflijk.
– Je ziet toch wel dat dat geen schaap is: 't is een ram, hij heeft horens ...
Nog een keer maakte ik mijn tekening overnieuw.
Maar die werd ook al geweigerd, net als de vorigen.
– Die is ook te oud. Ik wil een schaap, dat lang bij me blijft.
Toen werd ik ongeduldig, want ik wilde gauw beginnen mijn
motor uit elkaar te halen.
Ik maakte weer een krabbeltje voor hem en legde uit: "Dat is de kist. Je schaap zit erin."
Tot mijn verwondering zag ik de ogen van mijn kleine kunstcriticus stralen.

– Ja, zo wilde ik het helemaal hebben! – Denk je dat het schaap veel hooi nodig heeft?
– Waarom?
– Omdat ik maar een héél klein tuintje heb.
– Dat zal wel gaan. Ik heb je een héél klein schaapje gegeven.
Hij boog het hoofd over de tekening:
– Zo piepklein lijkt het nu ook weer niet ... Hé! Het is in slaap gevallen ...
En dat was dan mijn ontmoeting met de kleine prins.

(from Antoine de Saint-Exupéry, 2001/1943, chapter 2)

Samenvatting

Spraak is naast muziek het meest complexe akoestische signaal dat we regelmatig tegenkomen. Het signaal is rijk aan informatie die door een luisteraar geëxploiteerd kan worden om de door de spreker bedoelde boodschap te decoderen. Tegelijkertijd bevat het spraaksignaal non-linguïstisch informatie over de spreker, en is het vaak gemengd met andere geluiden uit de omgeving. Tot nu toe is het vermogen van het menselijk brein om een linguistische boodschap aan dit signaal te onttrekken niet geëvenaard door computers. Voor deze taak beschikken de hersenen over zeer gespecialiseerde systemen. Sommige hiervan zijn relatief statisch en zijn ontwikkeld in de loop van de evolutie of in de eerste maanden van het mensenleven, terwijl andere systemen dynamisch zijn en zich snel kunnen aanpassen aan een veranderende context. Het zijn de dynamische eigenschappen van het perceptuele systeem die het ons mogelijk maken om spraak zonder moeite te kunnen verstaan ondanks veranderingen van sprekers, accenten of achtergrondgeluiden — het soort factoren dat in het algemeen rampzalige consequenties heeft op de prestaties van computergestuurde spraakherkenningssystemen. Met dit proefschrift werd geprobeerd een bijdrage te leveren aan een beter begrip van de processen die deze snelle aanpassingen mogelijk maken. De focus ligt op het leerproces van luisteraars, die met een spreker geconfronteerd worden die een bepaalde spraakklank consistent op een ongewone manier articuleert. Meerdere thema's zijn betrokken bij dit onderzoek, zoals de relatie tussen spraakperceptie en de identiteit van een spreker, de wijze waarop snelle perceptuele aanpassingen gerelateerd zijn aan andere manieren van leren, hoe goed de huidige modellen van spraakherkenning dit proces kunnen verklaren, en welke neurale mechanismen hierbij betrokken zijn.

In alle experimenten die in dit proefschrift zijn beschreven werd het aanpassingsproces onderzocht door luisteraars te laten wennen aan spreker met een verschoven uitspraak van de klanken [f] of [s]. Luisteraars kregen herhaaldelijk een ambigue klank te horen in een specifiek woord. De luisteraars hoorden bijvoorbeeld het woord /kara?/ of /olij?/, waar [?] de ambigue klank representeert. Dit woord klonk alsof de spreker (de zogenaamde 'exposure spreker') zijn [f] meer als een [s] uitsprak. Nadat deze ambigue klank een aantal keren in het woord werd aangeboden, vertoonden de luisteraars een voorkeur om de klank [?] te interpreteren als een [f]. Een andere groep die herhaaldelijk dezelfde klank [?] in woorden zoals /naaldbo?/ of /radij?/ had gehoord, had de voorkeur om de ambigue klank als een [s] te interpreteren. Na blootstelling aan de specifieke uitspraak van een spreker kunnen luisteraars hun zogenaamde perceptuele grens tussen spraakklanken opschuiven in de richting van de uitspraak van de exposure spreker.

De experimenten in hoofdstuk 2 onderzochten of de verschuiving van de perceptuele grens specifiek wordt toegepast als luisteraars getest worden met spraak van de 'exposure spreker', die door door het produceren van ambigue klanken in lexicaal beperkte contexten de aanpassing veroorzaakte, of dat er generalisatie optreedt als luisteraars spraak horen van andere sprekers. Vanuit het perspectief van de luisteraar zou het efficiënt zijn om de aanpassing slechts één keer uit te voeren en deze vervolgens steeds te gebruiken wanneer naar spraak van de exposure spreker geluisterd wordt. Het is echter minder efficiënt om de aanpassing ongedifferentieerd toe te passen voor alle leden van de taalgemeenschap, tenminste als er geen aanleiding is dat anderen de kritische eigenschap gemeen hebben in hun spraakproductie.

De resultaten bevestigden inderdaad de sprekerspecifiekheid van deze vorm van perceptueel leren: Luisteraars pasten de verschuiving van de categoriale grens tussen [f] en [s] alleen toe op de fricatieven die werden geuit door de exposure spreker. Effecten van vergelijkbare grootte werden zelfs vastgesteld wanneer deze klanken werden gepresenteerd volgend op een klinker die werd uitgesproken door andere mannelijke en vrouwelijke sprekers, hetgeen een perceptie van sprekerverandering uitlokte. Geen effect werd gevonden voor testfricatieven die werden geproduceerd door een nieuwe spreker die de luisteraars nog niet hadden gehoord.

De in hoofdstuk 2 beschreven experimenten onderzochten verder of klank-intrinsieke informatie of klankextrinsieke informatie beïnvloed wordt door deze vorm van perceptueel leren. Klankintrinsieke informatie betreft alleen informatie over de spraakklank ([?]) zelf, terwijl klankextrinsieke informatie ook betrekking heeft op algemene kennis over de spreker, zoals de lengte en vorm van zijn of haar mond- neus- en keelholte. In Experiment 4 werden de ambigue klanken die tijdens de exposure- en testfasen worden aangeboden uitgesproken door een andere spreker dan de spreker die de lexicale exposure items geproduceerd had. Er werd een perceptueel leereffect vastgesteld dat net zo groot was als het effect in de condities met één spreker. Aangezien er tijdens de exposure fase geen andere informatie over de testspreker aanwezig was dan spektrale aanwijzingen van de ambigue segmenten, suggereren de resultaten dat dit type perceptuele aanpassing primair betrekking heeft op het verwerken van klank-intrinsieke aanwijzingen.

In hoofdstuk 3 werd onderzocht of een lexicaal-gestuurde aanpassing aan de foneemgrens een kortdurend verschijnsel is dat slechts voor korte tijd wordt onthouden en daarna wordt afgedaan, of dat deze aanpassing stabiel blijft over de tijd heen. Een groep luisteraars leerde in de ochtend om een ambigue klank als een [f] of als een [s] te interpreteren, en werd 12 uur later getest. Een tweede groep had de leerfase in de avond en werd de volgende ochtend getest. Beide groepen werden eveneens onmiddelijk na de exposure getest; op deze wijze werd een baseline meting gecreeërd. Twee factoren in dit design zouden in principe een stabieler effect kunnen produceren in de groep die de leerfase s'avonds had. Allereerst was het waarschijnlijk dat deze deelnemers in de tusentijd veel minder spraakinput van andere sprekers zouden ontvangen, waarin niet-ambigue producties van de kritieke spraakklanken voorkwamen die de gemanipuleerde aanpassing zouden kunnen corrigeren. Ten tweede hadden de deelnemers in de avondgroep tenminste zes uur geslapen voorafgaand aan de volgende testfase; zij hadden dus gelegenheid tot door slaap versterkte bestendiging van leren. Verondersteld wordt dat het gewend raken aan een ongewone uitspraak van een bepaalde klank een proces is dat luisteraars vaak gebruiken en dat dus zeer goed is aangeleerd. Om deze vorm van leren bruikbaar te maken voor de luisteraar, zou bestendiging niet veel tijd moeten vergen. De resultaten toonden aan dat er inderdaad een significant perceptueel leereffect optrad onmiddelijk na de training, en dat dit effect zowel in de ochtend- als in de avondgroep stabiel bleef. Er werden geen verschillen gevonden tussen de twee groepen in de grootte van het effect na een interval van 12 uur. Deze bevindingen suggereren dat perceptueel leren zeer stabiel blijft gedurende de testperiode, en dat er noch verval optreedt door interferentie van de spraak van andere sprekers in de wakkere uren, noch voordeel gehaald wordt uit de mogelijkheid tot bestendiging van leren tijdens de slaap.

Onderzoek naar de functionele neuro-anatomie van spraakperceptie heeft kandidaat-corticale gebieden geïdentificeerd waarin prelexicale verwerking van het spraaksignaal zou kunnen optreden. In samenspraak met psycholinguïstische modellen die een hiërachische organisatie in gesproken woordherkenning voorstellen, is er bewijs voor een mogelijke hiërarchische organisatie in de corticale gebieden die bij de auditieve woordherkenning betrokken zijn. Zo zijn gebieden die relatief specifiek op spraak reageren beschreven als lateral liggend ten opzichte van de primaire auditieve cortex, die betrokken is bij non-specifieke auditieve analyse. Vanuit deze gebieden zou informatie verder doorgegeven kunnen worden naar gebieden die meer vooraan op de temporaalkwab liggen, met name de voorste bovenste slaapwinding en -groeve.

Het experiment in hoofdstuk 4 maakte gebruik van functionele magnetische resonantie beeldvorming (fMRI) met een perceptueel leerparadigma. Het doel was verder te onderzoeken hoe prelexicale verwerking is geïmplementeerd in de neuroanatomie van de hersenen. Het uitgangspunt van het experiment was dat juist dat soort verandering in prelexicale verwerking die wordt geïnduceerd door perceptueel leren sterk bewijs zou kunnen leveren voor een functionele localisatie, indien correlaten hiervan gemeten zouden kunnen worden met fMRI. De voorspelling was aldus dat gebieden die gevoelig zijn voor een bepaald foneemcontrast verschillende reacties zouden laten zien wanneer dit contrast gemoduleerd wordt in de ene of in de andere richting door lexicaal gestuurd leren. Vier gebieden werden geïdentificeerd die overeen kwamen met het kritieke [f]/[s] contrast, in het bijzonder een gebied achter op de linker bovenste slaapwinding, hetgeen consistent was met verschillende andere studies die zich specifiek hebben gericht op prelexicale processen. Echter, het experiment vond in geen enkel gebied aanwijzingen voor een verandering in de neurale reactie als een functie van lexicaal-gestuurd leren. Dit negatieve resultaat roept vragen op van methodologische aard. Zo zijn in het bijzonder de keuze van een baseline conditie en het optimale niveau van ambiguïteit van de teststimuli relevante parameters die kunnen worden onderzocht in toekomstige onderzoek.

De experimenten in dit proefschrift en andere recente studies naar perceptueel leren in spraak hebben een mechanisme beschreven dat het perceptuele systeem in staat stelt zich snel aan te passen aan veranderende luistersituaties, waardoor perceptuele stabiliteit voor de luisteraar kan worden behouden. Door aan te tonen dat reeds in de hersenen opgeslagen lexicale representaties prelexicale verwerking over de tijd heen kunnen moduleren, en dat dit proces on-line analyse van de identiteit van de spreker vereist, hebben deze onderzoeken nieuwe beperkingen geïdentificeerd voor modellen voor gesproken woordherkenning. In de opbouw van toekomstige modellen zouden mechanismen voor de beïnvloeding van prelexicale verwerking – zowel door het sturen van leersignalen van het lexicon als door het aangrijpen vooraf ingewonnen spreker-specifieke informatie – moeten worden opgenomen.

Curriculum Vitae

Frank Eisner was born in 1976 in Saarbrücken, Germany. He studied Psychology and Communication at the University of Wales Institute, Cardiff, UK, and graduated in 2001. Later that year, he obtained a PhD stipend from the Max Planck Society for the Advancement of Science and joined the Comprehension Group at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. He is currently affiliated to the Speech Communication Group at the Institute of Cognitive Neuroscience, University College London, UK.

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 Schiller
3. Lexical access in the production of ellipsis and pronouns.
   Bernadette 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 Austronesian 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 categorization 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.
    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 electrophysiolgical 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: ERP studies on the time course in language comprehension.
    Oliver Müller
36. Lexically-guided perceptual learning in speech processing.
    Frank Eisner