Adjusting to different speakers:
Extrinsic normalization in vowel perception

ISBN: 9789076203416

Printed and bound by Ipskamp Drukkers b.v.

Adjusting to different speakers: Extrinsic normalization in vowel perception

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
volgens besluit van het college van decanen
in het openbaar te verdedigen
op maandag 28 november 2011 om 15.30 uur precies

door

Matthias Johannes Sjerps
geboren op 24 januari 1982
te Amsterdam

Promotoren:	Prof. dr. J. M. McQueen
	Prof. dr. A. Cutler
Copromotoren:	Dr. H. Mitterer (MPI)
Manuscriptcomissie:	Prof. dr. A. J. van Opstal
	Prof. dr. A. Van Wieringen (Katholieke Universiteit Leuven)
	Prof. dr. P. Boersma (Universiteit van Amsterdam)

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

Acknowledgements

Het volgen van een promotietraject is een fantastische uitdaging maar soms ook een zware opgave. Vaak prijs je jezelf gelukkig met de leukste baan op de wereld, maar op sommige momenten zit het vol tegenslagen en lijkt het enigszins nutteloos. De leuke momenten hebben er voor gezorgd dat ik zo ver ben gekomen dat ik straks mijn proefschrift mag verdedigen, de minder leuke momenten maken dat ik er ook trots op kan zijn. Er zijn een heleboel mensen die ik erg dankbaar ben. Deze mensen hebben, op allerlei verschillende manieren, bijgedragen aan wat ik de afgelopen jaren heb gedaan en hoe ik dat beleefd heb.

Ik denk dat ik toch moet beginnen bij mijn proefpersonen. Ik vond het zelf soms al een kwelling om mijn eigen experimentjes uit te zitten als ik ze ging testdraaien. Maar vrijwel al mijn proefpersonen hebben tientallen minuten naar ipapoes, epapoes, pitten en petten, soefoes en soofoos en "gekke kikker" of "zaag"-geluiden moeten luisteren. De data die ik aan hun ogen, oren, hersenen en vingers onttrokken heb was natuurlijk onmisbaar voor deze dissertatie. Al zullen ze dit waarschijnlijk nooit lezen, ik ben ze zeer dankbaar voor hun toewijding.

Eén van de mensen die het belangrijkst is geweest voor dit project, en mijn ontwikkeling als wetenschapper in het algemeen, is James McQueen. James, jouw rol begon al toen ik na mijn bachelor een folder las over de Master Cognitive Neuroscience. Jouw bijdrage daarin, over hoe mensen woorden herkennen in een stroom van "buzzes, bursts and chirps" raakte de juiste snaar en ik wist zeker dat ik deze master moest volgen. Na een succesvolle stage vroeg je mij of ik onder jouw begeleiding ook een promotietraject wilde volgen, waarop ik natuurlijk gretig ja heb geantwoord. Je bent snel van geest, creatief en kritisch, en daar heb ik veel aan gehad. Ik ben je dankbaar voor jouw bijna onuitputtelijke toewijding in het lezen, herlezen en herherlezen van mijn schrijfsels. Dat je meerdere malen dezelfde fouten moest verbeteren moet je vervelend hebben gevonden maar dat heb ik zelden aan je gemerkt. Ik heb me ook meerdere malen verbaasd over jouw capaciteit om resultaten in een compleet nieuw perspectief te kunnen plaatsen en er altijd positief naar te blijven kijken. Ik hoop dat ik me hier voor in de toekomst iets van eigen heb gemaakt.

De andere persoon die ik veel dank verschuldigd ben is Holger Mitterer. Holger, jouw passie voor het zo efficiënt mogelijk maken van allerlei dingen door te puzzelen met scripts en programmaatjes ben ik door jouw enthousiasme zowaar gaan delen. Ik kon altijd bij je aankloppen om mee te denken over allerlei praktische problemen of theoretische vragen. Je hebt verschillende projecten impulsen en nieuwe richtingen gegeven met je creatieve ideeën. Holger en James, voor promovendi in het algemeen is de verstandhouding met hun (co)promotoren soms een bijkomende bron van stress en ergernis. Ik kan oprecht zeggen dat ik me er altijd van bewust ben geweest dat ik veel geluk heb gehad. Onze meetings waren zinvol en ik heb er altijd veel plezier aan beleefd. Heel veel dank dus.

Natuurlijk ben ik ook Anne Cutler erg dankbaar. Anne, je hebt een groep wetenschappers bij elkaar gebracht die voor elke PhD een plek biedt om zich te ontwikkelen. Tijdens group-meetings was je altijd direct en eerlijk. Je bent een uitgesproken voorbeeld van een bevlogen wetenschapper, en dat is inspirerend. De Comprehension Group, en het MPI in het algemeen, vormden voor mij een fantastische plek om mijn PhD project te doen. Ik ben je dan ook dankbaar dat je me die mogelijkheid hebt geboden.

Met name wil ik natuurlijk ook mijn twee paranimfen, Susanne en Marijt, bedanken. Natuurlijk omdat jullie mijn paranimfen hebben willen zijn, maar vooral omdat jullie mijn tijd als PhD een stuk leuker hebben gemaakt. Susanne, voor veel van mijn proefpersonen was jij "de Stem". Dit was een opoffering, want ik was niet de enige op het instituut die jouw stem wilde gebruiken. Maar je was vooral een baken van gezelligheid en optimisme. Tijdens onze pauzes, fietstochtjes naar het station (en ja, naar Duitsland..... en terug..), en treinritjes heb je mij talloze keren op de hoogte gebracht van allerlei interessante nieuwtjes en dingetjes. Je was een fijn luisterend oor voor de onvermijdelijke problemen en frustraties die zijn langsgekomen. Marijt, we hebben vele uren tegenover elkaar gezeten, met twee schermen ertussen, tot het onvermijdelijke moment dat ik per ongeluk tegen een paar voeten aanschopte. Dat was natuurlijk meestal een welkome aanleiding om weer even over iets 'heel belangrijks' te kletsen. Ik moet hier natuurlijk ook bekennen dat ik dankbaar gebruik heb gemaakt van jouw mentale-encyclopedie-functie die de meest uiteenlopende informatie blijkt op te slaan. Laatst las ik dat het gebruik van Google er voor zorgt dat mensen minder onthouden. Wel, dan weet ik zeker dat er ook een onmiskenbaar

Marijt-als-kamergenoot effect te vinden moet zijn. We moeten maar zien of ik me nu alleen zal redden.

Verder zijn er natuurlijk een heleboel mensen die de afgelopen jaren een belangrijke plek innamen tijdens mijn leven op het MPI. Eva, ik kreeg pas gaandeweg door hoeveel parallellen er eigenlijk tussen onze projecten waren, maar die hebben uiteindelijk geleid tot een mooie samenwerking. Joost, in de periode met jou, Marijt en mij op een kamer steeg het niveau van flauwe en droge grappen tot een dieptepunt. Bedankt daarvoor. Caroline, we werkten tegelijk naar de eindstreep toe en konden daardoor veel frustraties maar ook successen delen! Verder heb ik leuke herinneringen aan een aantal anderen zoals Jiyoun (dank voor de sushi lessen), Annelie, Patrick, Katja, Wen Ciu, en een aantal van de eerdere student-assistenten zoals Laurence (die mij zelfs een tijdje gehuisvest heeft). Wat betreft mijn tijd in Nijmegen in het algemeen, en vooral tijdens de master die aan mijn promotie voorafging, heb ik veel plezier gemaakt met alle medestudenten uit mijn jaren van de CNS master.

Voor het interpreteren van data en het oefenen van praatjes heb ik veel gehad aan het commentaar van de verschillende mensen die bij de comprehension group-meetings aanwezig waren, zoals Mirjam, Mirjam, Esther, Alexandra, Falk (jaja binnenkort is het echt zover), Neil, Adriana, Agnieszka, en anderen. Verder zijn er een hoop mensen die me veel geholpen hebben met allerlei praktische zaken op het MPI. In het bijzonder Rian Zondervan, Jan Achterberg, Ad Verbunt, Alex Dukers, Johan Weustink en Pim en Thea.

During my PhD I visited the University of Texas at Austin for a couple of months. This was a wonderful and also very productive period. I am especially grateful to Rajka Smiljanić, because she gave me a warm welcome and immediately displayed a great deal of trust in me. Rajka, I felt very welcome in your lab, but also at your home for dinners. You are an ambitious scientist and such a nice and fun person. I am also very grateful to David Birdsong. David, you made sure that I had everything I needed when I came to Austin, this gave me a great start. As a result of this, and despite a fairly tight schedule, my project in Austin went very smoothly. I have many fun memories of our trips around the city and my time there in general. So thanks. I also received assistance from Anais Callejon, Rob Corona, Randall Rouse and Suzanne Baker. I thank Steven Miller for his efforts in dealing with visa and all sorts of impossible bureaucratic necessities. I had a number of fun evenings with Emilie and also with Zac, a perfect stranger at the Spiderhouse.

Inmiddels is mijn volgende baan alweer begonnen en daarom wil ik vooral Antje Meyer bedanken voor het vertrouwen. Het is heel leuk om in een nieuwe jonge groep te beginnen.

Naast mijn leven op het MPI bestond er natuurlijk ook een parallelle wereld waarin mijn proefschrift een haast mythisch concept was. De mensen die mijn leven maken tot wat het is, van ruim voor mijn PhD en zeker ook tot lang daarna. Ik hoop dat dit boekje een plek zal krijgen in jullie boekenkasten, al is het vergeeld en, op dit dankwoord na, grotendeels ongelezen.

Rijk, Erika, Mart, Kiki, Marije, en met tijd en wijlen aangevuld door Diederik, Roos en Just: als wij ergens afspreken, of zelfs op vakantie gaan, dan weten we steeds een knus wereldje te creëren waarin we leven als Bourgondiërs-boven-de-rivieren. Er wordt lekker gegeten, gedronken, gediscussieerd, gelachen, gedanst en gezongen (de laatste twee sporadisch, maar toch) en sinds kort is er zelf een accordeon! De ideale afleiding als ik even niet aan mijn proefschrift wilde denken.

Willem, Jantine, Nienke en Hans. Tsja, waar zal ik beginnen. Jantine en Nienke, ons gedeelde (studenten)leven begon op de Brem daarna aangevuld door Willem en Hans. Ik had me nooit beter kunnen wensen dan met jullie vier in een huis te wonen. Avondjes op de bank op de stoep zitten, achter buurtkinderen aanrennen om onze fietsventieltjes terug te halen en ga zo maar door. Jullie hebben mij altijd gesteund, ook al moest ik op gezellige avondjes doordeweeks soms al vroeg afhaken (gelukkig blijkt het nu dus niet voor niets!).

Ik wil ook Jan en Christina, en de hele familie Vandenbroucke en aanhang, bedanken dat ik zo welkom ben in de familie, bij etentjes en bij andere gezellige gelegenheden. We hebben de afgelopen jaren samen al veel bijzondere momenten meegemaakt.

In het bijzonder ben ik mijn familie dankbaar, voor allerlei dingen natuurlijk. Pa en Ma, ik wist dat jullie mij onvoorwaardelijk zouden steunen, wat er ook van mij terecht zou komen. Ik heb mij daardoor vrij gevoeld om uitdagingen aan te pakken, ook al wist ik van tevoren niet of ze te moeilijk zouden blijken om af te maken. Igor, Maria, Megan en Maarten. Jullie hebben ook altijd wonderbaarlijk veel vertrouwen in mij gehad, dank daarvoor. Ik ben vooral heel blij dat wij het als familie samen zo gezellig hebben. En het zal met kleine Mats erbij vast alleen nog maar beter worden! Heel veel dank allemaal.

Lieve Annelinde, het meest speciale bedankje gaat natuurlijk naar jou. Het is niet gelogen als ik zeg dat jij een van de belangrijkste rollen hebt gespeeld. Je was er altijd voor me als ik ergens mee zat, maar ook om het te vieren als ik een succesje had behaald. Ik weet niet waar ik het aan te danken heb, maar ik heb een leuke en mooie vriendin die ook nog eens weet hoe moeilijk het schrijven van een proefschrift soms kan zijn. Maar dat niet alleen, je begrijpt zowaar iets van de inhoud van mijn proefschrift en je hebt me vaak goede adviezen gegeven. Bedankt voor je steun, alle fijne momenten die we hebben gehad, en alle leuke dingen die we nog gaan doen.

Contents

INTRODUCTION

Chapter 1

Our ability to understand speech is a precious one. It allows us to communicate with others in a very fast, direct and sophisticated manner. For normal-hearing listeners it seems a fairly effortless way to receive information. In contrast to this subjective feeling, however, perceiving speech is a highly complicated operation. This complexity is exemplified by the fact that, despite the increase of computing power and the advancement of automatic speech recognition in general, machine speech recognition is still not up to par with human speech recognition ability (Benzeghiba et al., 2007). Moreover, human listeners themselves do not attain the ability to understand speech overnight. One of the main reasons for our difficulty with the recognition of speech is the fact that the speech signal is highly variable.

Variability: A formidable problem

The focus in this dissertation will be on variability of vowels and the ways in which listeners deal with this variability. Speech sounds, including vowels, are produced by the vocal tract. The timbre of vowels is a direct result of the movement of the vocal cords that provide a periodic sound source, the shape of the vocal tract and the position of the articulators (tongue, lips, etc.). The tongue often partitions the vocal tract into different sections that have their own resonance frequency. The resulting resonance properties of the vocal tract amplify or attenuate certain frequencies of the source signal. The auditory result of these combined factors is what listeners generally perceive as speech.

As a result of the amplification or attenuation of frequency regions, the speech signal contains frequency bands of relatively high amplitude. These are referred to as 'formants'. Formants are one of the main cues to identifying phonemes, especially vowels (Fry, Abramson, Eimas, & Liberman, 1962). For instance, the lowest formant (F_1) almost solely determines the difference between the vowels in the Dutch words "pit" (transcribed as /pɪt/, meaning: *the stone of a fruit*) and "pet" (transcribed as /pɛt/, meaning: *cap*). The mean values of F_1 for a female speaker of Dutch lie around 535 Hz for /ɛ/ and around 400 Hz for /ɪ/ (Adank, Smits, & van Hout, 2004). A difference

in frequency in F_1 of roughly 140 Hz can thus define two completely different phonemes. Together, F_1 and F_2 (and to some extent F_3) are, along with the vowel's duration, the main determinants of all vowels in Dutch.

So far, vowel perception sounds like a rather trivial task: The human speech recognition system should somehow measure the frequencies of the formants and this tells the listener which vowel was produced. The problem, however, is that the absolute values of formants vary a lot across situations and individuals. Therefore, a one-to-one mapping between absolute formant values and phonemes does not exist and hence absolute formant value estimation cannot straightforwardly result in correct vowel perception. In the following I will discuss three causes of such variance. Although this list is not exhaustive, it presents three of the main influences that are also the most relevant for this thesis.

The first cause of variance is the fact that speech signals are highly dependent on their immediate phonemic context (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). To exemplify, an instance of the Dutch vowel /ɑ/ (as in "*bak*") cut out from the non-word "rar" (transcribed as /rɑr/) will have very different formant values (and thus sound very different) than the same vowel cut out from the non-word "lal" (transcribed as /lɑl/). This is caused by the fact that the articulators are in slightly different positions when one produces the /l/ or the /r/. One way to look at this is that during the production of the vowel, the articulators are being pulled away from their "context-free" target positions (Nearey, 1989). These different places of articulation for the consonants have a strong effect on the way the intermediate vowel sounds. Figure 1 displays two recordings of my own voice while I pronounced "rar" (panels on the left) and "lal" (panels on the right). The top panels display the recorded signal as an amplitude waveform (oscillogram), and the bottom panels display its spectrogram. A spectrogram displays the same signal but then transformed so that the amount of energy in the different frequency regions becomes visible (blackness represents more energy). Different frequencies are plotted on the y-axis, time is represented along the x-axis. As one can observe in the spectrograms, both words contain two bands of energy between 500 and 1500 Hz. These are F_1 and F_2. The middle parts of the two non-words both contain the vowel /ɑ/ ("a"). The formants for this vowel, however, are not in exactly the same location.

A measurement in the middle of the vowel part showed that for "rar" F_1 is at 650 Hz, and F_2 is at 1162 Hz. For "lal" F_1 is at 623 Hz while F_2 is at 977 Hz. There is

thus a difference in the measured F_2 values of 185 Hz when comparing the /ɑ/ in /rar/ with that in /lal/. This shows that coarticulatory context has a strong influence on the way a vowel sounds.

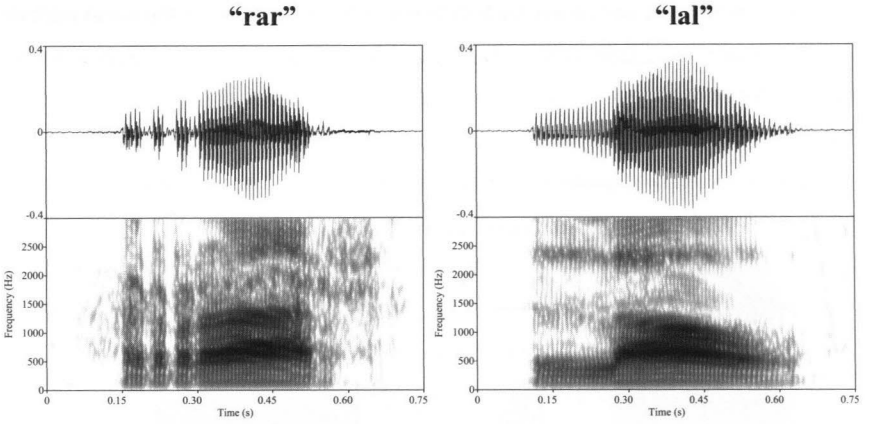

Figure 1. *The oscillograms (top panels) and spectrograms (bottom panels) of two Dutch nonsense words: "rar" (left panels) and "lal" (right panels).*

The second cause of variance is fact that every individual has their particular way of pronouncing phonemes. One could for instance think of differences between the way a word is pronounced by speakers from different regional backgrounds, speakers of different native languages that are producing words in the same non-native language (Flege, Munro, & Mackay, 1995), or even speakers with different kinds of speech impediment.

The third source of variation, which is the most relevant with respect to this thesis, are differences between speakers due to vocal tract properties. As stated above, the resonance properties of the vocal tract depend on its shape. An important way in which vocal tracts differ is in size. Not only does the length of the vocal tract increase when children get older and bigger, among adults the vocal tracts of males tend to be longer than those of females. This causes males to have, on average, lower formants. Figure 2 provides an example of this. It displays the productions of the Dutch word "kaak" (transcribed as /kak/, meaning *jaw*), spoken by a female (left panel) and a male (right panel) speaker. It appears that for the female speaker the F_1 and F_2 have higher values. The female speaker has an F_1 of 1118 Hz and an F_2 value of 1808 Hz. The male speaker has an F_1 of 819 Hz and an F_2 of 1615 Hz. Especially the difference in F_1 is quite large.

Even among speakers of the same sex, however, there are obvious differences in the physical characteristics of their vocal tracts. Apart from differences in the length of the vocal tract, speakers have different voices due to influences such as the position of their teeth and so on. Such differences in the makeup of the vocal tract all have important influences on the sounds that those vocal tracts produce (Peterson & Barney, 1952).

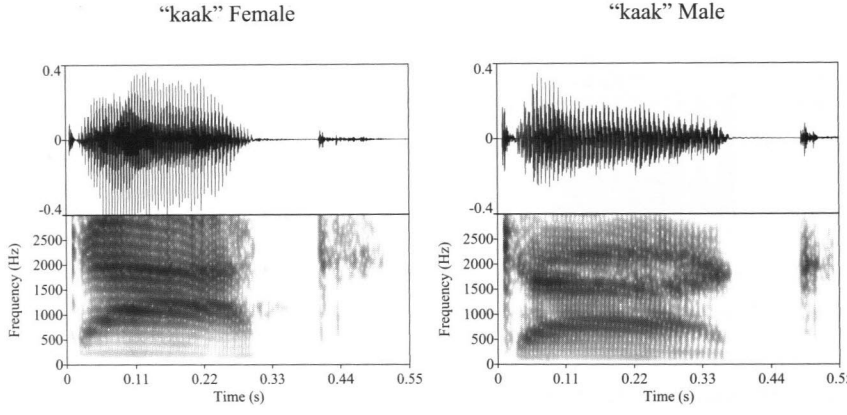

Figure 2. *The oscillograms (top panels) and spectrograms (bottom panels) of the word: "kaak". This word was produced by a female speaker (left panel) and a male speaker (right panel).*

Apart from these three main influences on variation, the way a phoneme is realized is also influenced by situational influences such as its place in the prosodic structure of the utterance (e.g., is it in stressed or unstressed position and uttered at the beginning or at the end of a sentence?), whether a speaker pronounces a word carefully or more colloquially (Johnson, 2004) and by speaking rate. Changes in production also arise due to phonological changes such as assimilation. For instance, in Dutch a fricative is devoiced after a voiceless obstruent: The voiced fricative /v/ in "zoutvat" (written as "v") is pronounced as /f/ because it occurs right after the voiceless obstruent /t/ (Booij, 1995). Furthermore, apart from variance due to the production of the speaker, listeners also have to deal with environmental sounds that might obscure the speech that they are trying to listen to (imagine listening to speech while riding a crowded bus).

The result of these combined influences is that the speech signal that reaches the listener's ear is variable and varies in many different ways. The different sources

of variability add up and thus result in relatively poorly defined categories. Those influences can be so large that, for instance, single instances of two different vowels, spoken by different speakers in different situations, can have very similar acoustic configurations. This means that depending on the situation, a single speech sound can be interpreted as one of a number of different phonemes. One can imagine that this poses a problem for listeners. In order to understand speech, speech sounds have to be mapped onto mental representations, that then combine to form meaningful units such as words. If the realization of phonemes varies, however, a one-to-one mapping between a sound that enters the ear and the mental representation of a phoneme category is difficult. This thesis deals with some of the ways in which listeners manage to deal with this problem.

The listener: A problem-solver with a sophisticated toolbox

Listeners appear to be relatively unaffected by the complicating influences of variability. It turns out that they are able to deal with the different forms of variability by virtue of a number of cognitive mechanisms. These mechanisms operate in different ways, but have a similar result: They allow the listener to map speech sounds onto their correct phonemic representations. Listeners are therefore mostly unaware of the fact that, when listening to speech, they are continuously solving a complex problem.

The categorical nature of phonemes

Phoneme category representations in listeners' brains are naturally sensitive to differences between phoneme categories. During speech perception the difference within phoneme categories is perceptually diminished while between-category differences are enhanced (Liberman, Harris, Hoffman, & Griffith, 1957). While listeners can in principle hear the differences within and between categories to a similar extent (Schouten, Gerrits, & van Hessen, 2003), the perceptual differences between phonemes become relatively larger when listeners focus on phonemic content. As an example of this finding, Iverson and Kuhl (1995) have shown that in a task where listeners have to indicate whether two subsequent vowels are the same or different (an AX, same/different discrimination task) they find it easier to detect that sounds are different when the two sounds belong to different categories than when they belong to the same category (although the physical distance between the two pairs is the same). This shows that listeners tend to disregard within-category variation. As mentioned earlier, however, in everyday speech the variability between

realizations of a particular phoneme category is so large that, without additional compensation processes, these sounds might still be recognized as the wrong phoneme categories. The tendency for listeners to be more sensitive to differences between phoneme categories than to differences within categories is thus not enough to resolve the variability problem.

Cross-modal influences

Another way in which listeners deal with the fact that the speech signal is variable is that they use multiple sources of information. One can imagine that when the auditory speech signal is ambiguous, due to for instance background noise, listeners can benefit from visual information. Listeners, however, also use visual information when the auditory signal is not degraded. In a classic experiment, McGurk and Macdonald (1976) showed that when listeners see and hear a speaker, they integrate the information from both sources. When listeners heard a speaker saying non-words like "pa-pa" and were asked to repeat what they had heard the speaker saying, they were mostly correct. However, when participants listened to the same speech while also watching a speaker who uttered different phonemes (e.g., "ka-ka"), listeners' responses were often incorrect even though, again, they were instructed to report what they *heard*. Their responses showed that they often formed a fused percept (i.e., auditory "pa-pa" with visual "ka-ka" could be perceived as "ta-ta"). This indicates that listeners use more than one of the available sources of information when listening to speech sounds.

Perceptual learning

Listeners are also able to deal with speaker-specific characteristics by rapidly adjusting their sound-to-phoneme mappings to the idiosyncrasies of a particular speaker. Imagine a speaker who naturally lisps. This speaker's realization of the phoneme /s/ is very different from realization of most other speakers for that category. Listeners, however, can rapidly adjust their phoneme category boundaries. For instance, when listeners are listening to a speaker who happens to produce either the /f/ or the /s/ as a sound that is ambiguous between these phonemes, then the boundary between /f/ and /s/ will shift to match this speaker's actual category boundary (Norris, McQueen, & Cutler, 2003). Perceptual learning is a powerful way to adjust one's perception to the specific situation at hand.

Compensation for coarticulation and perceptual context effects

Listeners also use information that is adjacent to the target speech sound (Lindblom & Studdert-Kennedy, 1967). As mentioned above, one of the causes of variability is the influence of neighboring phonemes. As another example, a phoneme /s/ will be realized with more lip-rounding in the context of /u/ than in the context of /a/. This can be problematic because in isolation a version of the phoneme /s/ with lip rounding sounds more like the phoneme /ʃ/ (the first phoneme in "shop"). Listeners compensate for this variance (Mann & Repp, 1980; Mitterer, 2006b). Mann and Repp (1980) asked listeners to categorize a continuum of sounds ranging from [ʃ] to [s]. Crucially, these sounds were followed by either the vowel "a" or "u" (for example, one of the targets was [su]). It was found that listeners categorized more sounds as the phoneme /s/ when it was followed by [u] than when it was followed by [a]. This shows that listeners compensate for the effects of coarticulation by shifting their category boundaries.

Speech rate normalization

Duration is another main determinant of phoneme identity. In Dutch, for instance, duration can indicate whether an instance of, for example, /s/ is part of a sequence of two instances of /s/ or only one. Consider the contrast between "eens speer" and "eens peer" ("once spear" vs. "once pear", note here that, contrary to intuition, there is no pause between the adjoining two instances of [s] in the first case: They are just produced as one longer [s]). In the first combination of words the [s] sound will be longer than in the second combination of words (in the second it only constitutes a single /s/). Listeners pick up on this durational cue and interpret the combination accordingly (Shatzman & McQueen, 2006). Again, however, variation in the realization of phonemes makes this distinction rather opaque. Consider two speakers, one speaking fast and the other speaking slow. When the fast speaker produces the double [s], it will have the same duration as the single [s] produced by the slow speaker. How can a listener then tell whether this speaker was saying "eens speer" or "eens peer"? Again, it has been found that listeners manage to compensate for this type of variation (Ainsworth, 1974; Summerfield, 1981). In a case like the example above, the perceived duration is normalized for the duration in the rest of the sentence (Reinisch, Jesse, & McQueen, 2011). This mechanism thus operates in a contrastive manner: if a sentence is spoken slowly (and hence long) then a subsequent sound will be perceived as shorter and if a sentence is spoken fast (and hence short)

then the perceptual mechanism will make the same subsequent speech sound sound relatively longer. This contrastive compensation mechanism thus partly resolves the variability introduced by differences in speaking rate.

Intrinsic normalization

The next mechanism makes use of the fact that speech signals not only contain cues about phoneme identity but at the same time also contain cues about more general properties of the speaker's vocal tract. These cues can help listeners to adjust the perceived signal to the overall properties of that speaker. To exemplify, the difference between /ɪ/ and /ɛ/ lies mainly in the first formant. As mentioned earlier, however, the height of the first formant is also determined by the length of a speaker's vocal tract. Crucially, the size of a speaker not only influences F_1, it also influences higher formants and it is (inversely) related to a speaker's average pitch. Intrinsic normalization is a mechanism that uses a combination of information from other formants (Sussman, 1986) and from the speaker's pitch (Miller, 1953) to estimate the speaker's general vocal tract properties. With respect to the example above (determining whether the vowel was /ɪ/ or /ɛ/), the first formant value will be interpreted relative to the estimated average vocal tract properties of the speaker. This mechanism is termed "intrinsic" because all the necessary information is available in the target phoneme (i.e., the formants and the speaker's pitch) and not in adjacent speech sounds.

The mechanism under investigation: Extrinsic normalization

The final mechanism, which will be the topic of this thesis, is extrinsic normalization. This mechanism is termed "extrinsic" normalization because, in contrast to the mechanism described in the previous paragraph, it relies on information *outside* the target vowel. In an influential experiment, Ladefoged and Broadbent (1957) showed that listeners compensate for vocal tract properties as revealed in a preceding sentence. They asked listeners to categorize a number of vowel sounds. These vowels were presented in the context of a carrier sentence that was manipulated to have either low or high vowel formants (thus simulating speakers with different vocal-tract shapes). Listeners were found to categorize a single instance of a vowel differently depending on the formant values in the carrier sentence. This shows that the formant values in a context sentence can influence the perception of a target sound. This effect has been observed on a number of occasions (Ainsworth, 1975; Broadbent, Ladefoged, & Lawrence, 1956; Mitterer, 2006a; Nearey, 1989; van

Bergem, Pols, & Beinum, 1988; Watkins, 1991; Watkins & Makin, 1994, 1996). It has also been found with naturally produced instead of manipulated speech (Ladefoged, 1989), with sine-wave versions of speech (Remez, Rubin, Nygaard, & Howell, 1987) and similar effects have been found for the perception of lexical tone (Francis, Ciocca, Wong, Leung, & Chu, 2006). There is a strong analogy between these findings and those showing normalization for speech rate and immediate coarticulatory context: listeners appear to perceive speech sounds in relation to their context, thereby reducing the problems caused by variability. In a way, this finding is not surprising if one thinks of other perceptual domains. Most sensory information is influenced by its context. For instance, have a look at the cover of this thesis. It can be observed that the person on the left appears to have a lighter color than the person on the right. In fact, these pictures have exactly the same brightness. It is the locally surrounding context that influences one's perception of the pictures. In the visual domain such effects are widely known (Hansen, Walter, & Gegenfurtner, 2007), but in the auditory domain such illusions also exist. Extrinsic vowel normalization is an example: the same ambiguous vowel can be interpreted in one of two ways as a function of the context in which it appears.

Extrinsic normalization: outline of the thesis

Relatively little is known about the details of when and how the brain actually normalizes for information in preceding context. Moreover, there are a number of findings that remain puzzling. For example, while recent findings suggest that normalization processes have an auditory basis, noise precursors can induce either strongly reduced normalization effects or no effects at all (Mitterer, 2006a; Watkins, 1991; Watkins & Makin, 1994, 1996). This thesis aims to increase understanding about the characteristics of the processes that cause extrinsic vowel normalization.

The first question that was posed was whether normalization is specific to the processing of speech sounds or whether it is a process that operates on sounds in general (Chapter 2). More specifically, it was investigated whether a signal's acoustic similarity to speech has an influence on the occurrence or amount of normalization that is observed. In this investigation, speech sounds were manipulated in a number of ways to make them more or less acoustically similar to speech. This investigation provides information about whether normalization is more likely to be the result of a general auditory process or of a process which applies only to speech.

The next chapter (Chapter 3) describes a set of experiments that investigate whether attention has an influence on normalization processes. If normalization is influenced by an attentional manipulation, this would show that the strength of the influence of a precursor on a subsequent target can be altered by higher-level processes. If not, however, this would suggest that normalization is for a large part driven by bottom-up information and thus mostly dependent on signal characteristics.

The next chapter describes research that was undertaken to investigate whether normalization is language specific (Chapter 4). This question was addressed in two ways: First, it was investigated whether listeners of different language backgrounds all normalize and whether they do so to an equal extent. Second, it was investigated whether listeners from those specific language backgrounds also normalize for speech spoken in other languages, and, if so, whether they normalize to the same extent across languages. These questions inform us about whether normalization is a mechanism that applies to linguistic input in general. One of the possibilities is that listeners who speak a language in which there is less between-category vowel overlap also normalize less. This would suggest that normalization is induced by the properties of the ambient language. In contrast, it could be that listeners normalize for preceding speech contexts to a similar extent, irrespective of whether the stimulus language is their second language or even a completely unfamiliar language. This would strengthen the conclusion that normalization is an auditory process that applies in a more general way.

Another important question about extrinsic normalization is whether it is task dependent. Previous experiments on extrinsic vowel normalization have used categorical tasks. In a categorical task, participants hear a single stimulus on every trial and are asked to choose between two response options, "A" or "B". Categorical tasks with speech stimuli encourage listeners to focus on speech-specific processing levels. In Chapter 5, a discrimination experiment was designed to test for normalization. In a discrimination task, listeners do not have to identify what they heard, instead, they indicate whether there was a difference between two or more stimuli. This task focuses listeners more on auditory aspects and less on categorical aspects of speech sounds (Gerrits & Schouten, 2004). This is especially so with the 4I-oddity task, in which the listener's task is to decide which stimulus in a sequence of four was the odd one out (AABA vs. ABAA: was the odd one out in second or third position?). If normalization effects are found only in a categorization task, this would

suggest that normalization processes are related to the listener's categorical representation of the speaker's vowel space. If, however, normalization effects are also found in a task in which listeners are encouraged to focus on auditory stimulus properties, this would suggest that normalization effects are due to a relatively low-level, auditory process.

The first five chapters seek to provide insights into the processing characteristics of extrinsic vowel normalization. In order to understand at what stage in the processing of speech extrinsic normalization operates, it is also important to find out at what point in time during processing it has an effect. Chapter 6 will present an electroencephalography (EEG) study which measured when the effects of the normalization process can be observed. If the EEG signals show that the consequences of normalization are observed only at a relatively late point in time during processing, then this would suggest that normalization is a relatively high-level process. If the consequences of normalization are observed during a relatively early time window, however, then this would suggest that normalization is caused by a relatively early and low-level process.

Finally, in the last part of this investigation (Chapter 7), it was tested whether there are hemispheric differences in normalization processes. It is known that the two hemispheres have different processing characteristics. For instance, speech processing is dominant in the left hemisphere. Differences in the amount and direction of compensation effects for stimuli that are processed more strongly in one or the other hemisphere could provide an explanation for why previous reports on normalization have sometimes provided very different results.

To summarize, this thesis investigates extrinsic vowel normalization – one of the important ways in which listeners deal with variation in the realization of phonemes. I approached this process from a number of different angles by using different methodologies such as categorization, discrimination, cross-language comparisons, EEG recordings and a task probing laterality of processing. This research was undertaken to obtain a better understanding of a fundamental process that affects our perception of speech in a continuous but unconscious manner.

Constraints on the processes responsible for the extrinsic normalization of vowels

---

Chapter 2

A version of this paper appeared as: Sjerps, M. J., Mitterer, H., and McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception & Psychophysics, 73*, 1195-1215.

Abstract

Listeners tune in to talkers' vowels through extrinsic normalization. We asked here whether this process could be based on compensation for the Long Term Average Spectrum (LTAS) of preceding sounds and whether the mechanisms responsible for normalization are indifferent to the nature of those sounds. If so, normalization should apply to non-speech stimuli. Previous findings were replicated with first formant (F_1) manipulations of speech. Targets on a [pɪt]-[pɛt] (low-high F_1) continuum were labeled as [pɪt] more after high- F_1 than after low- F_1 precursors. Spectrally-rotated non-speech versions of these materials produced similar normalization. None occurred, however, with non-speech stimuli that were less speech-like, even though precursor-target LTAS relations were equivalent to those used earlier. Additional experiments investigated the roles of pitch movement, amplitude variation, formant location, and the stimuli's perceived similarity to speech. It appears that normalization is not restricted to speech, but that the nature of the preceding sounds does matter. Extrinsic normalization of vowels is due at least in part to an auditory process which may require familiarity with the spectro-temporal characteristics of speech.

CHAPTER 2: CONSTRAINTS ON EXTRINSIC NORMALIZATION

Introduction

Our interpretation of auditory events is dependent on the context in which they occur. Context-dependent interpretation helps listeners resolve speech sound ambiguities such as those that arise from speaker differences. For example, the interpretation of vowels depends on the first and second formant characteristics of the speaker who utters those vowels (Ladefoged & Broadbent, 1957). One account of this process proposes that it is the result of a general auditory mechanism that normalizes perception of any acoustic input by constructing a long-term average of the distribution of frequencies of a sound source (Kiefte & Kluender, 2008; Watkins & Makin, 1996). Based on this long-term average, the perceptual impact of acoustic energy in certain frequency regions of a subsequent target sound becomes attenuated. This mechanism would thus influence the interpretation of speech sounds based on spectral information in the preceding sentence (or extrinsic context). According to this account, the same adjustments should apply to non-speech target sounds following non-speech precursors. The current paper investigates normalization of speech and versions of those stimuli that have undergone extensive manipulations to make them unlike speech. It was thus tested whether extrinsic normalization could be based solely on this general auditory Long Term Average Spectrum (LTAS) compensation mechanism. It was also tested whether compensation mechanisms can operate independently of the acoustic and perceptual characteristics of the precursor signal.

The key finding behind the LTAS mechanism is that suppressing or exciting particular frequency regions of a precursor sentence can induce a shift in the categorization of subsequent vowels. Watkins (1991) found effects of normalization when listeners categorized targets on an /ɪ/ to /ɛ/ continuum which were presented after intelligible precursor sentences that had been filtered. The filter suppressed either those frequencies that are generally more pronounced in an instance of /ɛ/ compared to /ɪ/ (i.e., an /ɪ/-minus-/ɛ/ filter), or the reverse (i.e., an /ɛ/-minus-/ɪ/ filter). Participants gave fewer /ɪ/ responses to targets after a precursor that was filtered with the /ɪ/-minus-/ɛ/ filter than after a precursor that was filtered with the /ɛ/-minus-/ɪ/ filter. Categorization appeared to be shifted because listeners were more sensitive to spectral properties that were suppressed in the precursor sentence, increasing the probability that the vowel was perceived as the one that had more energy in that frequency region. Watkins and Makin (1994, 1996) thus suggested that normalization

occurs through a process that could be described as applying an inverse form of the precursor's average filter characteristics to the target sound.

Similarly, Kiefte and Kluender (2008) tested participants' /i/ vs. /u/ categorizations on speech sounds from a 7x7 grid varying the second formant (F_2) and spectral tilt (both of these aspects can influence perceived /i/ vs. /u/ identity; Kiefte & Kluender, 2005). When a precursor sentence had been processed by the same acoustic filter that was used to adjust the spectral tilt of the target stimuli, listeners relied only on the target F_2 value. However, when the precursors had been filtered to match the F_2 peak of the following vowel, listeners' /i/ vs. /u/ responses were based on the target's spectral tilt alone. This effect shows that listeners suppress the information value of acoustic aspects that are invariant between the precursor and a subsequent target. The result is that listeners become less sensitive to static information, but gain sensitivity for acoustic change. Normalization for a signal's long-term frequency characteristics does just that.

Watkins and Makin (1994) found that normalization can also be observed with filtered precursor sentences that are played backwards, and that the amount of normalization is not reduced in this case. Furthermore, Stilp, Alexander, Kiefte and Kluender (2010) report normalization effects on the perception of musical instruments. They created a target range from "saxophone" to "French horn". Participants had to categorize these targets, which were presented after precursor sounds that were filtered to emphasize the spectral characteristics of either the French horn or the saxophone. They found – in analogy to previous findings with speech materials (e.g. Watkins & Makin, 1994, 1996) – that the categorization of the musical non-speech targets shifted depending on the spectral characteristics of the precursor signal. These shifts were also of a contrastive nature and were observed with speech and instrumental precursors.

The account emerging from this body of prior research is that extrinsic normalization in speech perception is based on a general-purpose auditory mechanism which compensates for the LTAS characteristics of preceding speech and is indifferent to the nature of the sounds from which the LTAS is derived. However, Watkins and Makin (1994) also found that the amount of normalization was strongly reduced when a noise precursor was used, even though it had the same LTAS as the speech precursor. Moreover, Watkins (1991) also found that the normalization effect can even be completely abolished when such noise precursors and the targets are

presented to opposite ears. This finding led them to argue that normalization effects take place at at least two different stages: First, an initial peripheral stage that only operates over short intervals and with ipsilateral presentation; second, a central normalization stage that operates over longer precursor-target intervals and applies to stimuli that are presented ipsilaterally and contralaterally. They argued that the small amount of normalization that was found with the noise precursor with bilateral presentation was therefore completely due to peripheral auditory compensation effects. Watkins and Makin (1996) also suggested that a prerequisite for normalization at central processing stages might be that a precursor signal needs to contain spectrotemporal variation.

Our focus in the present study was on normalization at central processing levels. We asked whether the central compensation mechanism is based solely on the LTAS, and hence whether it is completely indifferent to the exact nature of the precursor signal. The suggestion of Watkins and Makin (1996) that spectrotemporal variation is a prerequisite for normalization effects to occur at central processing levels already casts doubt on such a pure LTAS mechanism, but the process may have other prerequisites. There could be signals, acoustically intermediate between speech signals and signal-correlated noise, that also fail to induce central normalization effects. If so, this would raise the question why certain signals induce compensatory effects while others do not. In such a situation a learning account would potentially offer a solution. Because adult listeners have had an abundance of exposure to speech from different speakers, they will gain experience with the fact that certain voice properties are stable within a speaker. They could therefore learn that it is beneficial to perceive vowels relative to those voice properties. This can be achieved if listeners learn to normalize for LTAS properties of preceding sound sequences. Additionally, however, it would then be valuable for speech perception if central normalization was in some way restricted in order to avoid normalization for the wrong precursor signals. This means that it should only apply to signals with particular characteristics. We tested this learning hypothesis by manipulating the properties of precursor signals.

We began by testing whether compensation for the LTAS of a precursor signal is independent of the acoustic and perceptual characteristics of a precursor signal by comparing normalization at central processing levels in speech and non-speech signals. Non-speech signals were created by spectrally rotating speech sounds. Spectral rotation is a transformation that rotates the spectral make-up of a complex

signal around a central frequency, such that the information in the high frequency ranges trades places with the information in the low frequency ranges (Blesser, 1972). This transformation changes the frequencies of the formants but preserves the spectrotemporal complexity of the signal. Spectral rotation also keeps part of the pitch information of speech sounds intact through the repetition rate in the signal, as revealed by the pitch-synchronous amplitude envelope, and through the fact that the harmonics remain evenly spaced (Moore, 2003). Spectral rotation also inverts the spectral tilt. That is, while voiced speech sounds invariably have less energy at higher frequencies, spectrally rotated versions have more energy at higher frequencies. If normalization is specific to speech sounds, it should not occur with spectrally rotated versions of speech. The general auditory account, however, predicts that spectral rotation should not prevent normalization, because spectral rotation of both target and precursors keeps intact the overlap between the spectra.

In subsequent experiments we varied the auditory properties of speech and non-speech precursor signals. According to the general auditory account, both speech and non-speech signals should induce normalization effects through compensation for LTAS. According to an extension of the general auditory account that includes learning, the signal may need to have speech-like spectro-temporal characteristics, and thus compensation for LTAS would not be completely independent of more fine-grained temporal and spectral properties of the precursor signal.

Experiment 1

Experiment 1 tested the influence of speech precursors on speech sounds (Experiment 1a) and the influence of spectrally rotated precursors on spectrally rotated targets (Experiment 1b). In Experiment 1a, participants categorized targets on a [pɪt] to [pɛt] continuum (an F_1 distinction), presented after speech precursors with an increased or a decreased average F_1 level. In Experiment 1b, participants heard non-speech stimuli that were created by spectrally rotating the materials of Experiment 1a (both precursor and target stimuli).

As the participants in Experiment 1b were presented with novel non-speech stimuli, they first had to undergo a three-part training protocol to familiarize them with the materials. Participants in Experiment 1a underwent the same training with the speech materials to ensure that the amount and type of stimulus exposure was similar over the two experiments.

There are two potential problems using speech sounds and their spectrally rotated counterparts. The first stems from the fact that auditory frequency resolution decreases going from low frequencies to higher frequencies (Moore, 2003). Differences that would be audible in speech stimuli could become inaudible after spectral rotation. All materials were therefore low-pass filtered at 2.5 kHz, and spectrally rotated around a frequency of 1.25 kHz.

While this minimizes the differences in spectral resolution between the original and the spectrally rotated materials, participants may still find it easier to discriminate between extremely familiar speech sounds than between their unfamiliar spectrally rotated counterparts. The discriminability of the material sets was therefore equated via a pretest, in which participants heard the speech and non-speech stimuli in a staircase discrimination task (Appendix A). Using the results from this pretest, speech and non-speech stimuli were selected that differed by similar perceptual distances.

The second potential problem with speech and spectrally-rotated speech materials is that they may induce the type of peripheral normalization effects argued to be distinct from more central processes (Watkins, 1991). These effects, which operate over short interstimulus intervals, result in a shift in categorization functions in the same direction as normalization effects that take place at the more central levels of processing which are under investigation here. In order to prevent such peripheral effects with our materials, a precursor-target interval of 500 ms was used. This should thus ensure that any effect found is a true instance of longer-term normalization.

Experiment 1a, which involves speech materials, is expected to result in clear normalization effects. It should thus provide a replication of numerous previous studies (Ladefoged & Broadbent, 1957; Watkins, 1991; Watkins & Makin, 1994, 1996). This effect would be characterized by a shift in categorization functions for targets presented after a speaker with a generally high F_1 versus a speaker with a generally low F_1. The results for Experiment 1b, however, depend on the nature of the compensation process. No categorization shift should occur if extrinsic normalization is restricted to speech signals. But if normalization is a process that is not restricted to intelligible speech, we expect to observe a categorization shift similar to that predicted for Experiment 1a.

Method

Participants.

Participants of the Max Planck Institute for Psycholinguistics participant pool were recruited and tested until sixteen (eight for each part) had successfully completed the training and testing parts of the experiment (see Appendix B for details on all experiments and sub-experiments). They were all native speakers of Dutch, reported no hearing or language disorders, and had not participated in a similar experiment before. They received payment for their participation.

Materials.

Experiment 1a Targets.

All recordings were made by a female native speaker of Dutch. The materials were down-sampled offline to 11050 Hz. Acoustic processing of the stimuli was carried out using Praat software (Boersma & Weenink, 2005). The test syllables were the Dutch words /pɪt/ (the stone of a fruit) and /pɛt/ (cap). To create a test continuum, the vocalic portion of a recording of the word /pɛt/ was excised. Using a Linear Predictive Coding (LPC) procedure, the source model (a model of the sound emitted from the vocal folds) was separated from the filter model (a model of the filter characteristics of the vocal tract) using 20 predictors. Using fewer predictors left remnants of the formants in the source model, which would have made it more difficult to shift the perceived identity of the targets towards /ɪ/. The formant filter model was based on 4 formants. The continuum was created by a linear decrease of F_1 over 200 Hz in steps of 1 Hz in the formant model. The formant and filter model were recombined to create the target vowel continuum. The average F_1 value of the endpoint [ɛ] was 575 Hz, the average F_2 value was 1844 Hz (F_2 was not manipulated). The F_1 values are close to the average F_1 values found in female speakers of northern standard Dutch, which is closest to the dialect of the speaker (/ɛ/ = 535 Hz; /ɪ/ = 399 Hz; values from Adank, van Hout & Smits, 2004) while the F_2 value is relatively low (/ɛ/ = 1990 Hz; /ɪ/ = 2276 Hz). The manipulated vocalic portions were spliced back into the unmanipulated consonant context from [pɛt]. All materials were band-pass filtered between 200 and 2500 Hz. All targets were adjusted so that their overall amplitude and their amplitude envelope matched those of the original vowel instance of /pɛt/.

Based on the pretest (Appendix A), six target steps from the initial continuum were selected ranging from [pɛt] to [pɪt]. These steps spanned an F_1 range of 60 Hz in

steps of 12 Hz (with F_1 values ranging from 70 to 130 Hz lower than the recorded /pɛt/). The pretest showed that this frequency range resulted in clearly discriminable endpoints.

Experiment 1a Precursors.

The precursor was the Dutch sentence "op dat boek staat niet de naam" [ɔp dɑt buk stat nit də nam] (on that book, it doesn't say the name) which contains, among others, the vowels [u], [a], [i] and [ə], thereby providing listeners with all of the point-vowels and schwa. The precursor did not contain the target vowels /ɪ/ and /ɛ/, to prevent direct precursor-target comparison. The average F_1 value over the vocalic portions of the selected precursor sentence was 502 Hz, ranging from roughly 300 Hz (in [i]) to 800 Hz (in [a]). The F_1 values of the vocalic portions of the precursors were, in two versions, either increased or decreased by 200 Hz (after Watkins & Makin, 1994) using the same method as that which was used for the target vowels. The formant filter model was based on 4 formants except for the vowel portion of the word /nit/, which was based on 3 formants. Along with the surrounding vowels, the first nasal of the word /nam/ was also included in the manipulation as that increased the naturalness of the token. The other two nasals were unmanipulated in both conditions. The manipulated vocalic portions were spliced back into the unmanipulated consonantal parts of the original precursor sentence.

Experiment 1b Targets.

The targets were created in the same way as those for Experiment 1a, with the addition of the critical manipulation that the signals were spectrally rotated around 1250 Hz. The pretest (Appendix A) determined that the 60 Hz F_1 difference used for the speech targets was too small to be detected for the spectrally rotated versions of the targets. The F_1 difference between the endpoints for Experiment 1b was therefore set at the full range of 200 Hz, leading to approximately equally discriminable test stimuli across the experiments. A six-step continuum between the two endpoints was selected (steps of 40 Hz).

Experiment 1b Precursors.

The precursors that were used in Experiment 1a were spectrally rotated around 1250 Hz (the same manipulation as was applied to the targets for Experiment 1b).

LTAS measures.

Figures 1 and 2 display the LTAS (bin width = 10 Hz) of each precursor and each endpoint target, along with each difference LTAS, for both sub-experiments. The x-axis is logarithmic. Figure 1 shows that the Low F_1 precursor has more energy than the high F_1 precursor at low frequencies. Although for example the Low F_1 precursor LTAS is not perfectly matched to the LTAS of the targets, the difference lines show that the relative differences between the Low F_1 and the High F_1 stimuli are indeed matched to the differences between the target endpoints. This means that those frequencies that listeners use to distinguish between /ɛ/ and /ɪ/ are roughly the same frequencies that constitute the acoustic difference between the precursors in the High and Low F_1 conditions. Both precursors have more energy at higher frequencies than the targets (the target spectral tilt has a steeper slope). This should thus in principle cause an ambiguous target to be perceptually slightly shifted towards /pɪt/ for both precursor conditions. However, the difference line between the precursors shows that the induced shifts towards /pɪt/ will be stronger for the High F_1 condition as it has more energy at the higher frequencies than the Low F_1 precursor. The two precursor conditions are therefore predicted to induce different target categorization functions. The focus is therefore on the relative differences among the precursors matched to the relative difference among the targets[1].

The comparison between Figure 1 and 2 shows that the relation between the difference spectra for the precursors and their respective endpoint targets was similar across the two sub-experiments. The high- F_1 speech precursor had more energy at frequencies above the average F_1 value than the low- F_1 speech precursor, and the /pɛt/ endpoint had more energy at frequencies above the average F_1 frequency than the /pɪt/ endpoint. Spectral rotation reversed these differences but preserved the similarities in the relation between the precursors and targets. The spectrally rotated

[1] This discussion raises the question whether it would be possible to model the predicted direction and amount of normalization that should be observed with any combination of precursor and target. While this is beyond the scope of the current paper it would in principle be possible. A model would have to include a number of parameters. First, one would have to specify which frequency components are perceptually important for the distinction between two different vowels and to what extent. Second, a model would probably have to restrict the influence of any precursor frequency-component to only target frequency-components that have overlapping tuning curves. The size of these tuning curves will be dependent on the level of processing that is modeled (peripheral or central) as these might not be of an equal size. Third, as the contextual influence of a certain acoustic event is probably restricted to some amount of time, it is likely that the most recent part of a precursor signal has more influence on the direction and amount of compensation than parts of a precursor signal that are most distant (and also here the amount differs depending on the level processing that is modeled). This thus would require additional weighting functions for temporal distance.

precursor based on the high- F_1 speech precursor thus had more energy at frequencies below the average spectrally-rotated F_1 than the low- F_1 spectrally rotated precursor. This was also the case for the spectrally rotated /pɛt/ endpoint relative to the spectrally rotated /pɪt/ endpoint.

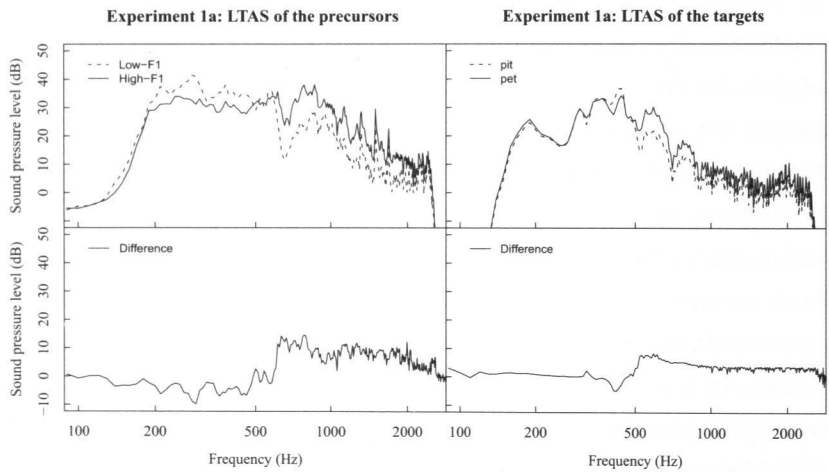

Figure 1. LTAS plots for the speech materials in Experiment 1a. Upper left panel: LTAS for the high-F_1 precursor (solid line) and for the low- F_1 precursor (dotted line). Bottom left panel: the difference spectrum for the precursors. Upper right panel: LTAS for each endpoint target, /pɛt/ (solid line) and /pɪt/ (dotted line). Bottom right panel: the difference spectrum for the targets.

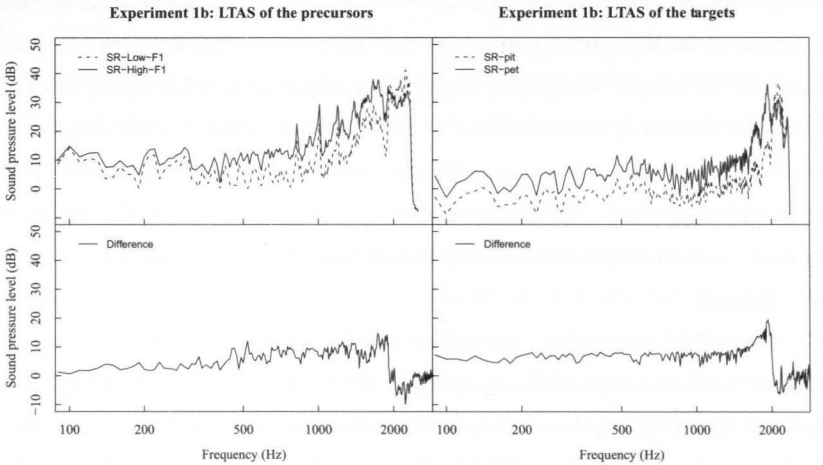

Figure 2. LTAS plots for the spectrally rotated materials in Experiment 1b. Upper left panel: LTAS of the spectrally rotated (SR) versions of the precursors with a high- F_1 (solid line) and for the low- F_1 (dotted line). Bottom left panel: The difference spectrum for the precursors. Upper right panel: LTAS for each spectrally rotated endpoint target /pɛt/ (solid line) or /pɪt/ (dotted line). Bottom right panel: the difference spectrum for the targets.

Design and Procedure.

The experiment was run using Presentation software (Version 11.3, Neurobehavioural Systems Inc.). All auditory stimuli were presented binaurally, through Sennheiser HD 280-13 headphones.

Training.

In the three-part training procedure, used in both sub-experiments, participants learnt to categorize the two unambiguous endpoint stimuli. The first part consisted of a discrimination experiment. On every trial participants heard a combination of the two endpoints (words in Experiment 1a; spectrally rotated versions of these words in Experiment 1b). The task was to indicate by pressing a button whether the two stimuli heard on a trial were the same or different. Visual feedback ("correct" (correct) or "fout" (incorrect)) appeared on a computer screen after each trial. If participants had seven out of eight responses correct on three consecutive blocks, they entered the second part of the training. For this second part participants listened to the same two endpoint stimuli, but in isolation. They were told that it was their task to find out which stimulus belonged to which button label ("A" or "B") using the feedback they

would receive, and that they would thus initially have to guess. Visual feedback was the same as in the first part. If participants reached a 90% correct criterion over three blocks (of 10 stimuli each) they entered the third part. This again involved categorizing these same two sounds as "A" or "B" with feedback, but the sounds were now presented after an unmanipulated (neutral) version of the precursor that was to be used in the testing phase. The same criterion applied as in the second part. Within all three parts, participants were allowed a self-paced pause after every hundred trials.

Testing.

In each sub-experiment, the six target steps were each played after the two precursors for fifteen repetitions, resulting in 180 test trials (with two pauses). This phase took about 12 minutes. Trials were presented without feedback. Participants categorized the targets by means of the same two buttons used during the second and third training phases ("A" and "B").

Data analysis.

In this and all following experiments, responses faster than 100 ms after target onset were excluded. Luce (1986) shows that simple responses to auditory stimuli start from 100 to 150 ms after stimulus onset. Any faster responses could thus not have been due to the perception of the target stimuli. After exclusion of missed responses and responses that were too fast, 99.7% of the trials were kept on average over Experiments 1, 2, 3 and 4 (the lowest proportion of preserved responses was 99.5%; no fast responses needed to be excluded in Experiment 5). The data were analyzed using linear mixed-effects models in R (version 2.6.2, R development core team, 2008, with the lmer function from the lme4 package of Bates & Sarkar, 2007). Different models were tested in a backward elimination procedure, starting from a complete model. All factors were entered as numerical variables, centered around 0. These include the factors Step (level on the continuum; -2.5 to 2.5 in steps of 1), Precursor (levels; Low F_1 (-1) vs. High F_1 (1)), Block (15 stimulus repetitions; levels -7 to 7 in steps of 1), and all their possible interactions. Non-significant predictors were taken out of the analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was also found. The best model was then established by means of a likelihood ratio test.

Results

Experiment 1a.

The upper panel of Figure 3 shows the average categorization results. The optimal model for the data had significant main effects of the factor Step ($b = -0.777$, $p < 0.001$), which indicates the decrease of /pɪt/ responses towards the /pɛt/ end of the continuum, Block ($b = -0.030$, $p = 0.046$), which indicates that the number of /pɪt/ responses decreased as the experiment progressed, and Precursor ($b = 0.556$, $p < 0.001$). The latter effect indicates that the probability of a /pɪt/ (low F_1) response is much higher after High than after Low F_1 precursors. Interaction effects were found between the factors Step and Block ($b = 0.033$, $p = 0.001$), indicating that participants' categorizations became, overall, less categorical as the experiment progressed, and between the factors Block and Precursor (b -0.03984, $p = 0.010$), reflecting the fact that the effect of Precursor became smaller towards the end of the experiment, although it was never absent.

Experiment 1b.

The bottom panel of Figure 3 displays the results. The optimal model had main effects of the factor Step ($b = -0.726$, $p < 0.001$), which indicates that the probability of a "non-speech /pɪt/" response decreases when moving along the continuum from "non-speech /pɪt/" to "non-speech /pɛt/", and an effect of the factor Precursor ($b = 0.196$, $p = 0.002$). The effect of Precursor indicates that the probability of a "non-speech /pɪt/" (NS-/pɪt/) response is higher for targets after a spectrally-rotated high F_1 precursor than for those after a spectrally-rotated low F_1 precursor. This is an effect in the same direction as for the speech stimuli.

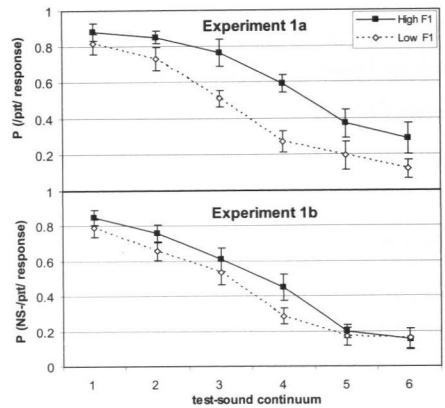

Figure 3. Experiments 1a (upper panel) and 1b (lower panel): Probability of a /pɪt/ or "Non-speech /pɪt/" (NS-/pɪt/) response to a continuum of the targets /pɪt/ (step 1) and /pɛt/ (step 6) in Experiment 1a or the spectrally rotated analogs of these in Experiment 1b. Targets were presented after precursor sentences with an F_1 value that was increased (high- F_1) or decreased (low- F_1) by 200 Hz (or the spectrally rotated analogs for Experiment 1b). Error bars reflect standard errors.

Discussion

Experiment 1a showed that the spectral properties of a precursor sentence can influence the categorization of a vowel continuum. This replicated earlier findings (e.g., Ladefoged & Broadbent, 1957; Watkins & Makin, 1994). Experiment 1b showed that this finding is not restricted to the processing of speech sounds. Spectrally rotated versions of the stimuli used in Experiment 1a resulted in a normalization effect that was similar to that found in Experiment 1a. The size of this effect was reduced in Experiment 1b, but this probably reflects the fact that the difference in frequency between the endpoint stimuli was much larger for the spectrally rotated stimuli (60Hz for speech, 200Hz for the spectrally rotated speech). The step size for the spectrally rotated targets in Experiment 1b was much larger (steps of 40 Hz) than the step size for the speech targets in Experiment 1a (steps of 12 Hz). Watkins and Makin (1996) have demonstrated that the ratio of the spectral contrast over the target continuum to that of the precursor continuum has a strong influence on the size of the categorization shifts. With respect to our stimuli, the difference spectrum for the targets of Experiment 1a was smaller than the difference spectrum of the targets in Experiment 1b whereas the difference spectra for the

precursors in Experiment 1a and Experiment 1b were of the same size. Given the findings by Watkins and Makin (1996), this could have resulted in the smaller normalization effect in Experiment 1b. As is evident from the results of the pretest, a larger target step size was necessary for participants to categorize the non-speech target range reliably. The focus here is thus on the qualitative finding that normalization occurred with unintelligible stimuli, and over a relatively long precursor-target interval.

Two additional observations from the results of Experiment 1a deserve to be discussed. The first is that in Experiment 1a the effect of Precursor decreased as the experiment progressed. This is probably due to the fact that the effect of a precursor potentially extends beyond the trial in which it is presented. As the different precursor conditions were presented intermixed, listeners could have been increasingly influenced by precursors from both conditions. It is unclear why this effect did not occur in Experiment 1b however. The second observation is that the proportion of /pɪt/ responses decreased as the experiment progressed. It is unclear what the exact nature of this effect is, but it is possible that listeners started the test phase with a slight bias towards the /pɪt/ end of the continuum and compensated for this bias as the experiment progressed (Repp & Liberman, 1987).

In sum, however, the results of Experiment 1 support the LTAS compensation account of extrinsic normalization. Moreover, the auditory compensation mechanism seems to apply to both speech and non-speech materials suggesting that the central processing mechanism is general, and independent of the acoustic and perceptual aspects of the precursor (apart from its LTAS). This conclusion seems to conflict with the finding by Watkins (1991) that with noise precursors no effects of precursors are found. One might therefore argue that the materials in Experiment 1b were too much like speech, and hence that the results could be explained by a speech-specific mechanism. Although spectral rotation destroyed the phonetic content of the original sentence, the precursors used in Experiment 1b still had many speech-like prosodic characteristics. In Experiment 2, the materials were manipulated in ways that still preserved spectrotemporal variation in the precursors. More importantly, it also preserved the LTAS relation between the precursors and the targets. However, the manipulations rendered the materials acoustically more unlike speech (compared to the materials in Experiment 1b). These manipulations consisted of: removing pitch variation, removing the (very) low amplitude parts (e.g., low-amplitude parts of stop

closures), temporally reversing the individual syllables, equalizing the average amplitudes of the syllables, and, again, spectral rotation of all materials. If normalization at a central processing level is the result of a general auditory process that compensates for LTAS, completely independent of the acoustic nature of the precursor, these extremely non-speechlike materials should still induce a normalization effect. If Experiment 2 does not result in a normalization effect, however, this would suggest that there are in fact restrictions to the type of precursor sounds that can induce normalization.

Experiment 2

Method

Participants.

Participants of the Max Planck Institute for Psycholinguistics participant pool were recruited and tested until eight had successfully completed the training and testing parts of the experiment (see Appendix B). Participants received a monetary reward for their participation. None had participated in Experiment 1

Materials.

Targets.

The targets were the same as those in Experiment 1b. But to maintain similarity to the new precursors, they were now manipulated such that they had a flat pitch level (at the average level of the original target). These pitch adjustments were made using the overlap-add method for resynthesis in Praat.

Precursors.

The precursors that had been used in Experiment 1a (Low F_1, unmanipulated, High F_1) were first modified to have a flat pitch at the average value of the speech materials (223.8 Hz) using the same method as was used for the targets. Each of these signals was then divided in high and low amplitude parts (see the upper panel of Figure 4: non-annotated sections are considered low amplitude). All the high amplitude parts were temporally reversed (e.g., the first digital sample of a word became the last sample of the new "reversed word" and vice versa), and equalized in amplitude relative to each other. Reversing only the individual syllables rather than reversing the complete sentence has the advantage that the pattern of LTAS-change as the sentence develops is very similar across the speech and non-speech stimuli (and much more similar than would be the case if the complete sentence were reversed). All low amplitude parts were excised and discarded. The resulting signals were then

spectrally rotated. The bottom panel of Figure 4 displays the resulting precursor and target in one condition. The manipulations that were applied in this and the following experiments are summarized in Table 1.

Table 1: Summary of Precursor Manipulations and Results

	Precursor manipulation					
	Normalization effect found					
Experiment	Pitch flattening	Reversed syllables	Spectral rotation	Equal amplitude	Breaks removed	
1a	No	No	No	No	No	Yes
1b	No	No	Yes	No	No	Yes
2	Yes	Yes	Yes	Yes	Yes	No
4a	No	Yes	Yes	Yes	Yes	No
4b	Yes	Yes	Yes	Yes	No	No
4c & 4d	Yes	Yes	No	Yes	Yes	Yes

Note. Precursor sentences were or were not manipulated with respect to several speech characteristics: Pitch movement, reversal of syllables, spectral rotation, syllables of equal amplitude, presence versus absence of low amplitude parts (breaks).

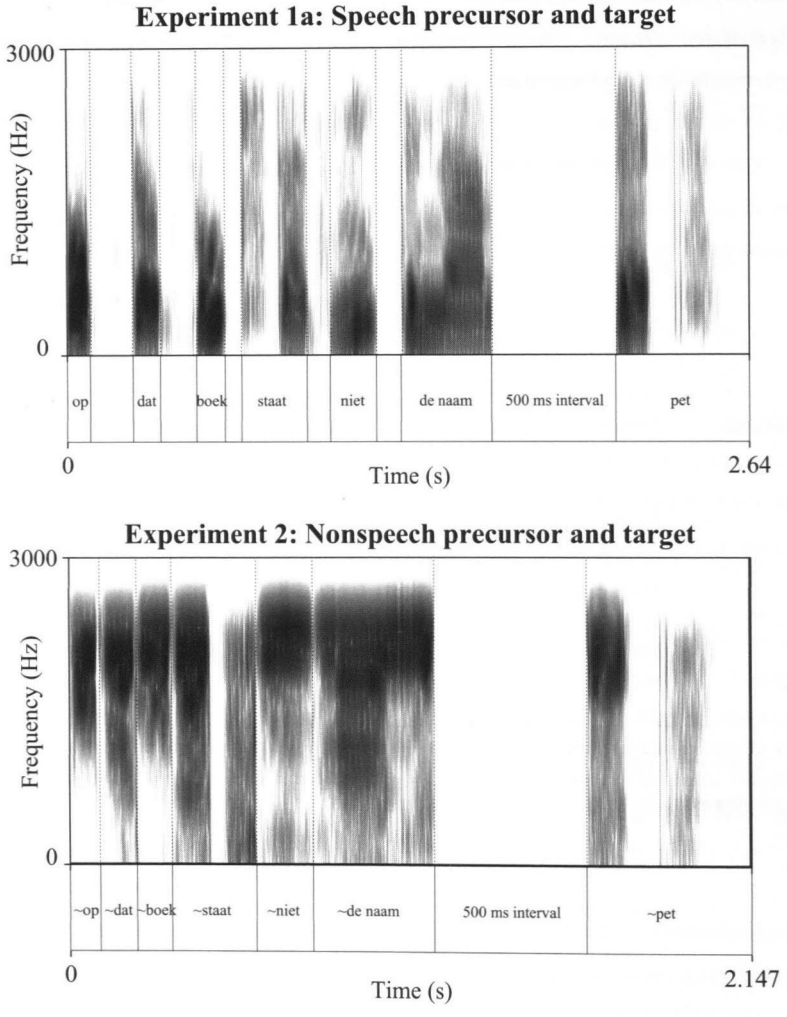

Figure 4. Upper panel: Experiment 1a. Annotated spectrogram of the original precursor sentence in the Low F_1 condition followed by /pɛt/. Bottom panel: Experiment 2. Annotated spectrogram of the manipulated precursor in the Low F_1 condition followed by non-speech /pɛt/. The symbol "~" indicates that the materials were manipulated.

LTAS measures.

Figure 5 displays the LTAS plots of the precursors and the endpoint targets, along with their difference LTAS. If compared to the LTAS plots from Experiment 1b (see Figure 2), the current LTAS plots may seem quite different at first glance, with reasonably smooth spectra in Figure 2 and peaky spectra in Figure 5. This is a consequence of the pitch flattening procedure; the peaks represent the harmonics of the constant f0. If one focuses on the difference between the spectra of "High- F_1" and "Low- F_1" versions, however, the relations are similar across Experiments 1 and 2. The non-speech "High- F_1" precursor had more energy at frequencies below the average F_1 value than the "Low- F_1" precursor. Similarly, the "non-speech /pɛt/" endpoint had more energy at frequencies below the average F_1 frequency than the "non-speech /pɪt/" endpoint.

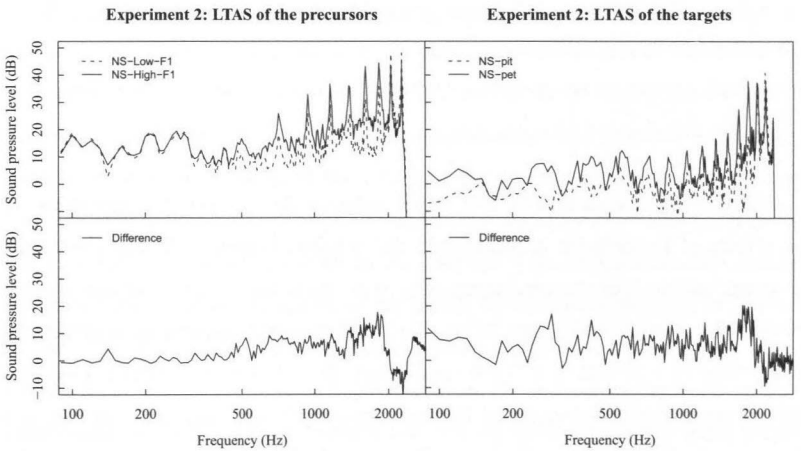

Figure 5. LTAS plots for the non-speech materials in Experiment 2. Upper left panels: LTAS for the non-speech sounds that originated from precursors with a high- F_1 (solid line) and a low- F_1 (dotted line). Left bottom panels: The difference spectrum for the non-speech precursors. Upper right panels: LTAS for each non-speech endpoint target that originated from the speech sounds /pɛt/ (solid line) or /pɪt/ (dotted line). Right bottom panels: the difference spectrum for the non-speech targets.

Results

Figure 6 displays the results. Modeling settled on main effects for the factors Step ($b = -0.945$, $p < 0.001$), indicating robust categorization, and Block ($b = -0.057$, $p<0.001$), which reflects a drift toward less overall "non-speech /pɪt/" (NS-/pɪt/) responses during the experiment. An interaction was found between the factors Step

and Block (b = 0.040, p < 0.001), which shows that responses became less categorical as the experiment progressed. When the factor Precursor was included, it did not have a significant effect (b = -0.116, p = 0.084). There was only a trend in the direction opposite from that observed in Experiments 1a and 1b, and this trend was limited to one step.

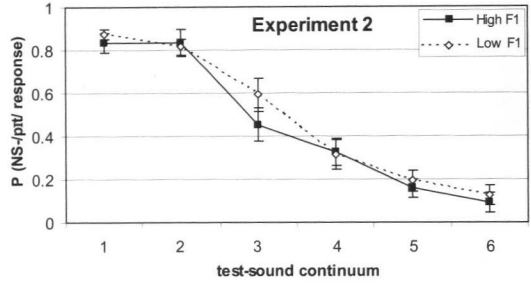

Figure 6. Experiment 2: Probability of "non-speech /pɪt/" (NS-/pɪt/) responses to a continuum of non-speech targets that were manipulated versions of a range from /pɪt/ (step 1) to /pɛt/ (step 6), presented after manipulated versions of the precursor sentences used in Experiment 1a. Error bars reflect standard errors.

Additionally a comparison was made between the effects in Experiment 1b and the effects of Experiment 2, to compare the effects obtained with these two non-speech experiments. The factor Experiment was included in the analysis (levels; Experiment 1b (-1) vs. Experiment 2 (1)). The optimal model showed main effects for the factors Step (b = -0.836, p < 0.001) and Block (b = -0.025, p = 0.019). Two-way interactions were found between the factors Experiment and Step (b = -0.104, p = 0.001), responses in Experiment 2 were more categorical than those in Experiment 1b, Experiment and Block (b = -0.032, p = 0.003), there was a decrease in the number of "non-speech /pɪt/" only in Experiment 2, and Experiment and Precursor (b = -0.157, p < 0.001). The latter effect reflects the critical comparison between the effects of Precursor over Experiments 1b and 2. The effect shows that there was significantly more normalization with the materials in Experiment 1b. A three-way interaction was found between the factors Experiment, Step and Block (b = 0.029, p < 0.001), indicating that only in Experiment 2 did responses become less categorical as the experiment progressed.

Discussion

In contrast to the normalization effects obtained in Experiments 1a and 1b, Experiment 2 failed to show a normalization effect for non-speech sounds. A comparison between the experiments showed that the size of the effect of the context condition was statistically different. An absence of compensation for LTAS was found in Experiment 2 despite a precursor-target LTAS relation that was very similar to that in Experiment 1b. As a comparison between the bottom panels of Figures 2 and 5 shows, the precursor difference-LTAS overlaps with the target difference-LTAS for both experiments in a similar way. If extrinsic normalization would have been the result of compensation for precursor-LTAS, independent of the exact nature of the precursor, it should also have been found in Experiment 2.

It thus appears that a general LTAS compensation account that assumes indifference to the nature of the precursor signal is inadequate for centrally located compensation processes. However, one could argue that the presence of normalization in Experiment 1b but not in Experiment 2 is due to acoustic differences between the stimuli. It is possible that the acoustic manipulations for Experiment 2 led to some subtle change in the relation of the LTAS of the precursors and targets, not apparent from the difference spectra, which prevented the normalization effect. To investigate this possibility a control experiment was set up. This experiment was designed to establish what the effect would be of the LTAS of the precursors used in Experiments 1b and 2 if it was physically applied to their respective target stimuli. The target sounds from each experiment (Experiment 3a for the Experiment 1b stimuli, and Experiment 3b for those from Experiment 2) were filtered such that those frequencies that were most pronounced in the precursor sentences would be most suppressed in the target signals. This method physically influences the target sounds in the same way as the hypothesized mechanism for LTAS compensation (c.f. Watkins 1991), in analogy to the approach taken in Watkins and Makin (1994). As the precursors of Experiment 1b elicited a significant shift in target categorization, it was predicted that Experiment 3a would result in a strong shift in categorization. If the lack of a categorization shift in Experiment 2 was due to insufficient overlap between the difference LTAS of the precursors and targets, Experiment 3b should not show a shift in categorization. If, however, there is sufficient overlap between the difference LTAS of the Experiment 2 precursors and targets, we should find a shift in categorization in Experiment 3b that is similar to that predicted in Experiment 3a.

Experiment 3

Method

Participants.

Eight further participants of the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. They received a monetary reward for their participation.

Materials.

The training materials in Experiments 3a and 3b were the endpoint stimuli that were used as training and test materials in Experiments 1b and 2 respectively. The test materials in Experiments 3a and 3b were filtered versions of the test materials used in Experiments 1b and 2. To create these stimuli the LTAS of each precursor from the earlier experiment was applied as an inverse filter to all steps of the target continua. This means that the amplitude for every frequency was attenuated by the relative average amplitude of the precursor signal at that frequency. Those frequencies that were most pronounced in the precursor signals were thus also relatively most suppressed in the new target sounds. This operation mimics the situation that perception of the frequency distribution of a target sound is perceived relative to the frequency distribution of the precursor to its full extent. As the manipulation resulted in signals with very low overall amplitudes, the amplitudes of all targets were increased by 20 dB (equally across the whole spectrum) such that participants would be able to listen to these new targets at a comfortable hearing level. Figure 7 displays the LTAS of the resulting "non-speech /pɛt/" target sounds filtered by both the High F_1 and the Low F_1 precursor from Experiment 2 (the LTAS for only one step of the continuum is displayed because the difference spectra are the same for all steps). A comparison of the resulting difference spectrum with the difference spectrum of the precursors from Experiment 2 (Figure 5, left bottom panel) shows that they are indeed the same. This means that we successfully applied the inverse LTAS filters of the precursors to the target continua. The same manipulation was applied to the targets of Experiment 1b (filtered by their appropriate precursor).

The same participants took part in both sub-experiments. Half the participants were trained and tested on the Experiment 3a materials first, and then on the Experiment 3b materials; for the other participants this order was reversed. Training in the first sub-experiment (3a or 3b) consisted of the first two phases from the earlier experiments (discrimination, and then categorization) using the identical endpoint

stimuli as were used in either Experiment 1b (for Experiment 3a) or Experiment 2 (for Experiment 3b). The first testing phase consisted only of the new precursor-LTAS filtered target continua, coming from the appropriate earlier experiment. In the second sub-experiment, only the second training phase (categorization) was presented (the stimuli differed only in pitch characteristics, and thus were already familiar to the participants). This second training phase was followed by the second test phase (again, with the appropriate precursor-LTAS attenuated target continua).

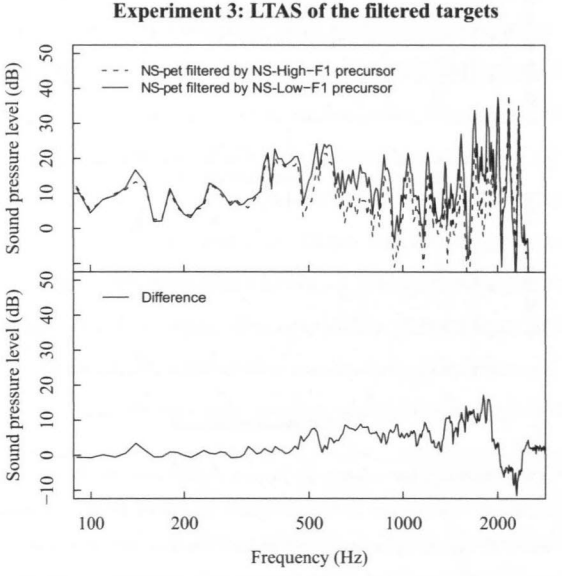

Figure 7. Upper panel: LTAS plots for the endpoint "non-speech /pɛɪ/" targets from Experiment 2, attenuated by a filter that was constructed from the low- F_1 precursor (solid line) and the high- F_1 precursor (dotted line) taken from Experiment 2. Bottom panel: The difference spectrum for the resulting two Experiment 3 sounds. Whereas for the other figures the difference spectrum were calculated by subtracting the "Low- F_1" variant from the "High- F_1" variant, here the subtraction was made in the opposite direction for ease of comparison.

Results

Experiment 3a.

The top panel of Figure 8 displays the results. The order of the two parts was added as the factor Order (levels; first: (-1) vs. second (1)) in the data analysis. The optimal model showed main effects for the factors Precursor-filter, i.e., which

precursor from the earlier experiment was used to filter the target, (b = 2.307, p < 0.001) and Step (b = -0.879, p < 0.001). Two-way interactions were found between the factors Step and Order (b = 0.117, p = 0.009, indicating fewer categorical responses if this part was presented as the second one) and Order and Block (b = -0.056, p = 0.003, indicating fewer "non-speech /pɪt/" responses towards the end of this part of the experiment).

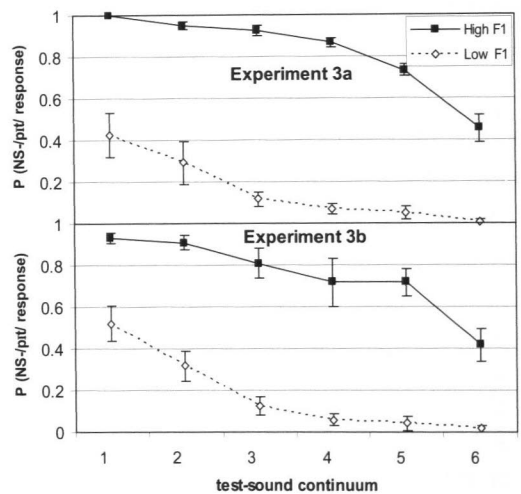

Figure 8. Control experiments, Experiments 3a (upper panel) and 3b (bottom panel). The graphs displays the probability of "non-speech /pɪt/" (NS-/pɪt/) responses to the continuum of non-speech targets from Experiment 1b (upper panel) or Experiment 2 (bottom panel), ranging from "non-speech /pɪt/" (step 1) to "non-speech /pɛt/" (step 6). The targets were attenuated by the filter properties of the LTAS of the non-speech High-F_1 or Low-F_1 precursors that were used in Experiment 1b (upper panel) or Experiment 2 (bottom panel). Error bars reflect standard errors.

Experiment 3b.

The bottom panel of Figure 8 displays the results. The optimal model showed main effects for the factors Precursor-filter (b = 1.856, p < 0.001) and Step (b = -0.762, p < 0.001). Two-way interactions were found between the factors Precursor-filter and Step (b = 0.158, p = 0.005, indicating a stronger effect of Precursor-filter towards the "non-speech /pɛt/" end of the continuum) and Precursor-filter and Order (b = -0.281, p < 0.001, indicating a smaller effect of Precursor-filter if this part was presented as the second one).

Experiment 3a vs. 3b.

An additional comparison was made between the effects obtained in the two sub-experiments. This analysis included the factor Experiment, that modeled the difference between the two sub-experiments (levels; 3a: (-1) vs. 3b (1)). The optimal model showed an effect for Precursor-filter ($b = 2.096$, $p < 0.001$), Step ($b = -0.833$, $p < 0.001$) and Experiment ($b = -0.215$, $p = 0.004$, indicating less overall "non-speech /pɪt/" responses in Experiment 3b). Two-way interactions were found between Experiment and Precursor-filter ($b = -0.215$, $p = 0.005$, indicating a smaller effect for Precursor-filter in Experiment 3b), Precursor-filter and Order ($b = -0.221$, $p = 0.004$, indicating a smaller effect of Precursor-filter if this sub-experiment was presented second) and Step and Order ($b = 0.112$, $p = 0.008$, indicating less categorical responses if this part was presented as the second one). Three-way interactions were found between the factors Precursor-filter, Step and Experiment ($b = 0.117$, $p = 0.006$, reflecting the fact that the interaction between Precursor-filter and Step was not present in Experiment 3a), Precursor-filter, Block and Order ($b = -0.035$, $p = 0.005$, indicating a progressive decrease in the effect of Precursor-filter but only if this sub-experiment was presented second), Precursor-filter, Experiment and Order ($b = -0.168$, $p = 0.002$, reflecting that the difference in size of the effect of Precursor-filter across Experiment 3a and 3b was only present when comparing sub-experiments that were presented second).

Discussion

Experiments 3a and 3b both resulted in large categorization shifts. The spectral relation between the precursors and targets was thus similar across Experiments 1 and 2, that is, these control experiments demonstrate that the LTAS relation between the precursors and targets of both earlier experiments is such that a compensatory influence from the precursors should result in a contrast effect in both cases. The lack of a normalization effect in Experiment 2 is thus apparently not due to an insufficient match of spectral properties between the precursors and targets of those materials. It appears instead that the properties of the precursors in Experiment 2 were not appropriate for central normalization to take place.

It should be noted that the compensation effect in Experiment 3a (derived from the spectrally rotated speech of Experiment 1b) was bigger than the compensation effect in Experiment 3b (derived from the more extreme non-speech materials of Experiment 2) and this was only the case when comparing sub-

experiments that were presented second. While we have no straightforward interpretation of this pattern, it is unlikely that this small difference in the size of the effect in this control experiment could explain the large difference between the results of Experiments 1b (compensation effect) and Experiment 2 (no compensation effect, with a trend in the opposite direction).

As Experiment 2 showed that not all precursor signals influence perception of subsequent targets in the same way, a LTAS compensation process that is indifferent to the exact nature of the precursors cannot fully account for whether normalization effects occur. This raises the question which type of precursors give rise to compensation for LTAS. There are a number of differences between the precursors used in Experiment 1b (which elicited compensation effects) and Experiment 2 (which failed to elicit compensation effects). Any one of these acoustic aspects might account for the absence of effects in Experiment 2. Experiment 4 addressed this issue by testing whether the presence of pitch variation in the precursor signals is a crucial factor (Experiment 4a), whether the presence of high and low amplitude parts in the signal can induce normalization processes (Experiment 4b); and whether a speech-like spectral tilt plays a critical role (Experiments 4c and 4d). In natural speech, these aspects are, to some degree, almost always present in the signal. In the materials for Experiment 2, however, they were removed. If any of them are necessary for extrinsic normalization to occur, one or more sub-experiments should reveal a normalization effect.

Experiment 4

Method

Participants.

For each of the four sub-experiments, eight different participants were recruited from the Max Planck Institute for Psycholinguistics participant pool (see Appendix B). Participants received a monetary reward for their participation. None had taken part in earlier experiments.

Materials.

For Experiments 4a and 4b, the targets were identical to those used in Experiment 2 (spectrally rotated speech with a flat pitch contour). The precursors of Experiment 4a were created by imposing a sinusoid pitch contour (formula = 223.8 + 79.1 * sin [t* (3.14 / (1.78.2))]) onto the precursors of Experiment 2. The contour had two periods and a pitch range of a similar size to the range of the pitch contour in the

speech precursors. The overall amplitude was set at the same level as for the other experiments. For Experiment 4b, the materials were created by reintroducing zero amplitude parts into the precursors of Experiment 2 at the same locations as the low amplitude parts in the speech version (Figure 4a). However, these parts were all of equal duration (0.107 ms), which was the average duration of the low amplitude parts in the original speech materials. The materials of Experiment 4c (both precursors and targets) consisted of those from Experiment 2, but then spectrally rotated back, such that low frequencies became high frequencies and vice-versa. The result of this operation is that the formants reappeared in their original locations, and the signal regained a speech-like spectral tilt.

It would however be interesting to compare the results of Experiment 1a (speech that had all the speechlike characteristics) with the results of Experiment 4c (speech that was manipulated such that some of the speechlike characteristics were removed). This would be difficult, however, as the target range was different between Experiments 1a and 4c. To allow for this comparison an additional experiment was run. Experiment 4d used the same precursor as Experiment 4c but then with the target sounds from Experiment 1a that had an F_1 range of 60 Hz.

Results

Experiment 4a: Pitch movement.

The top left panel of Figure 9 displays the average categorization results. The model that optimally explained the data consisted of a single effect for the factor Step ($b = -0.941$, $p < 0.001$). If the factor Precursor was included, it did not show a significant effect ($b = 0.051$, $p = 0.452$).

Experiment 4b: Low amplitude parts.

The results are shown in the bottom left panel of Figure 9. The optimal model consisted of an effect for the Intercept ($b = -0.392$, $p = 0.046$), which indicates that the probability of a "non-speech /pɪt/" response was smaller than 0.5, and an effect for Step ($b = -1.166$, $p < 0.001$). If the factor Precursor was included, it did not result in a significant effect ($b = -0.062$, $p = 0.394$).

Experiment 4c: All but spectral rotation (200 Hz F_1 range).

The top right panel of Figure 9 displays the average categorization results. The optimal model showed a main effect for the factors Step ($b = -2.110$, $p < 0.001$), Block ($b = -0.068$, $p = 0.002$) and Precursor ($b = 0.247$, $p = 0.010$). The precursor effect reflects a small but significant effect in the predicted direction. Additionally, a

three-way interaction between the factors Step, Precursor and Block was found (b = 0.070, p = 0.003). This reflected the fact that in the first half of the experiment the effect of precursor was more pronounced on one side of the continuum whereas it was more pronounced on the other end of the continuum in the second half of the experiment.

Experiment 4d: All but spectral rotation (60 Hz F_1 range).

The bottom right panel of Figure 9 displays the results. The optimal model showed a main effect for the factors Step (b = -1.050, p < 0.001) and Precursor (b = 0.137, p = 0.0498). The latter reflects a very small but significant effect in the predicted direction. A two-way interaction was found between the factors Step and Block (b = -0.025, p = 0.030), indicating that responses became more categorical as the experiment progressed.

Additionally a comparison was made between the effects in Experiment 1a and the effects of Experiment 4d, including the factor Experiment (levels; Experiment 1a (-1) vs. Experiment 4d (1)). The optimal model showed main effects for the factors Step (b = -0.908, p < 0.001) and Precursor (b = 0.338, p < 0.001). Interactions were found between the factors Experiment and Step (b = -0.140, p < 0.001) and Experiment and Precursor (b = -0.202, p < 0.001). The latter shows that the effect of precursor is strongly reduced in Experiment 4d (when the precursors lacked speechlike prosodic characteristics). A three-way interaction was found between the factors Experiment, Step and Block (b = -0.026, p < 0.001), which reflects that in Experiment 1a participants' responses became increasingly less categorical whereas in Experiment 4d responses became increasingly categorical as the experiment progressed.

Discussion

Experiments 4a and 4b show that reintroducing pitch movement or silent periods in the signal does not change the influence of a precursor sentence on subsequent targets. Spectrally rotating the materials of Experiment 2 again (such that the formant structures appeared in their original locations and the signal had a speech-like spectral tilt) resulted in small but significant normalization effects in Experiments 4c and 4d. Note that Experiment 1b and Experiment 4c are in a way opposites. The manipulation that was applied to the stimuli of Experiment 1b (spectral rotation) was the only manipulation that was not applied to the materials of Experiment 4c. It thus appears that a signal with a speechlike spectral tilt and formant structures can induce a

normalization effect, but that this is not the only signal characteristic that can do so. The combined speech-like acoustic aspects that were available in the materials of Experiment 1b led to a numerically larger effect than the speech-appropriate formants in Experiments 4c and 4d.

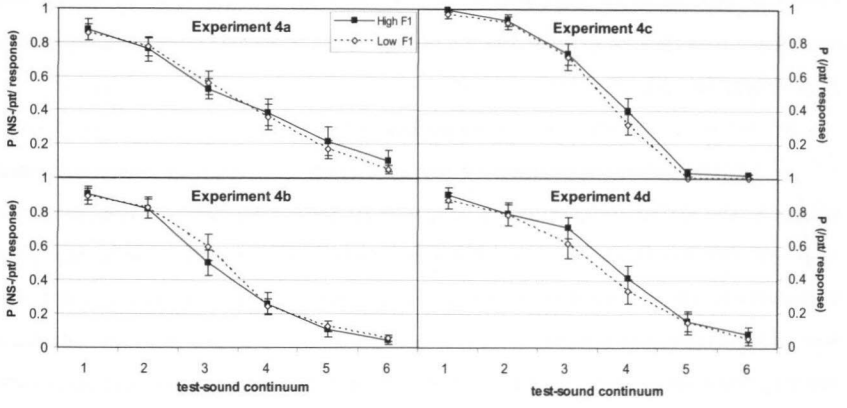

Figure 9. Experiment 4. Upper left panel: Results for Experiment 4a. The graph displays probabilities of "non-speech /pɪt/" (NS-/pɪt/) responses to the continuum of non-speech targets ranging from "non-speech /pɪt/" (step 1) to "non-speech /pɛt/" (step 6). Targets were presented after non-speech precursors which were created from both the High-F_1 and Low-F_1 non-speech sounds of Experiment 2 which now did have pitch movement. Bottom left panel: Results for Experiment 4b. The graph displays the same probabilities as those in the top left panel. Targets were presented after non-speech precursors which were created from both the High-F_1 and Low-F_1 non-speech sounds of Experiment 2 which now had low amplitude parts. Upper right panel: Results for Experiment 4c. The graph displays probabilities of /pɪt/ responses to the continuum of targets ranging from /pɪt/ (step 1) to /pɛt/ (step 6). Targets were presented after precursors which were created from both the High-F_1 and Low-F_1 precursors of Experiment 2 which now had been spectrally rotated back. The targets' F_1 frequencies ranged over 200 Hz. The result of this was that the formant structures appeared in their original positions. Error bars reflect standard errors. Bottom right panel: Results for Experiment 4d. The graph displays the same probabilities as those in the top right panel. The precursor was the same as in Experiment 4c. The targets were the same as those used in Experiment 1a, and thus covered an F_1 range of 60 Hz. Error bars reflect standard errors

The comparison between Experiments 1a and 4d showed that the influence of a precursor was attenuated when the precursor lacked a number of speechlike prosodic characteristics. It thus appears that the speech-like acoustic aspects that were available in the materials of Experiment 1b and the speech-appropriate formants in the materials of Experiment 4c and 4d induced similar compensation processes. In speech materials like those of Experiment 1a, all these speech-like acoustic aspects are available. This would explain the relatively larger normalization effect found in Experiment 1a.

Experiment 5

An alternative explanation of the occurrence (or absence) of normalization effects in the different experiments presented so far is that such differences are the result of differences in how speechlike the stimuli were perceived. The stimuli in Experiments 1b and 4c and 4d (where normalization was found) were not only acoustically similar to speech, but could also have been perceptually more similar to speech, while those in Experiments 2 and 4a and 4b (where no normalization was found) may have been both acoustically and perceptually less similar to speech. This would suggest that a listener's overt perception of the "speechiness" of the materials could induce a compensatory strategy, resulting in normalization effects only for those materials that listeners judge to be similar to speech. Although informal discussion with the participants after the experiments does not support this interpretation, it is a possibility that deserves more direct investigation. Experiment 5 was set up to investigate this matter by asking a new group of participants to rate the different precursor signals that were used in this series of experiments on their similarity to speech. If perceived speechiness influences the amount of normalization, it would be expected that the precursors of Experiments 1b and 4c and 4d (which elicited normalization effects) would be judged to sound more speechlike than the materials in Experiments 2, 4a and 4b. A lack of these patterns would argue against a role of speechiness in the induction of normalization effects.

Method

Participants.

Sixteen further participants of the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. They received a monetary reward for their participation.

Materials.

Eight types of precursors were presented. Five consisted of the same precursors that were used for Experiments 1b, 2, 4a, 4b, and 4c and 4d. These five types were each presented in their Low F_1, Neutral (i.e., as used in the third part of the training phases of the experiments), and High F_1 versions, resulting in 15 different precursors. Additionally, participants also rated three more types of sounds: Noise, unmanipulated speech, and band-pass filtered speech. The noise precursors consisted of three versions: noise with the same amplitude envelope as the precursors from Experiment 1; noise with the same long-term average spectrum as the neutral precursor from Experiment 1; and a combination of these two manipulations. The unmanipulated speech consisted of three sentences, spoken by the same speaker, and recorded during the same recording session as the materials of the previous experiments. The band-pass filtered speech consisted of filtered versions (200-2500 Hz, as for the materials in the previous experiments) of these unmanipulated speech sentences. This resulted in a total of 24 (5x3 for the precursors from the previous experiments and 3x3 for the additional types of carriers) different precursors. The precursors were randomly presented three times, during consecutive blocks. The unmanipulated speech and the noise conditions provided perceptual anchors for the other conditions.

Participants were asked to indicate how "speech-like" they thought the stimuli were. Participants heard a precursor once, and then made their judgment by moving a vertical cursor along a (51-step) horizontal bar with on the left of the bar "niet-spraak" (non-speech), and on the right "spraak" (speech). Participants moved the cursor using the mouse wheel, which started in the middle of the bar on every trial.

Results and Discussion

Figure 10 displays the average "speechiness" ratings. The precursors that were most extremely manipulated (those used in Experiment 2) were not rated as more or less speech-like than those of Experiments 4a (non-speech with pitch movement, $b = -0.618$, $p = 0.429$) and 4b (non-speech with low-amplitude parts, $b = -0.222$, $p = 0.775$), but they were rated less speech-like than those of Experiment 4c and 4d (all non-speech manipulations except for spectral rotation, $b = 10.70$, $p < 0.001$). The precursors of Experiment 2 were rated as slightly more speech-like than the precursors of Experiment 1b (non-speech with only spectral rotation, $b = 1.736$, $p < 0.043$). The latter result suggests that there is no relation between perceived

speechiness and amount of normalization, as such an account would predict that the precursor used in the experiment with the most extremely manipulated stimuli (Experiment 2) would be judged considerably less like speech than the precursors used in the non-speech experiment with only spectral rotation (Experiment 1b), because only the latter elicited normalization effects.

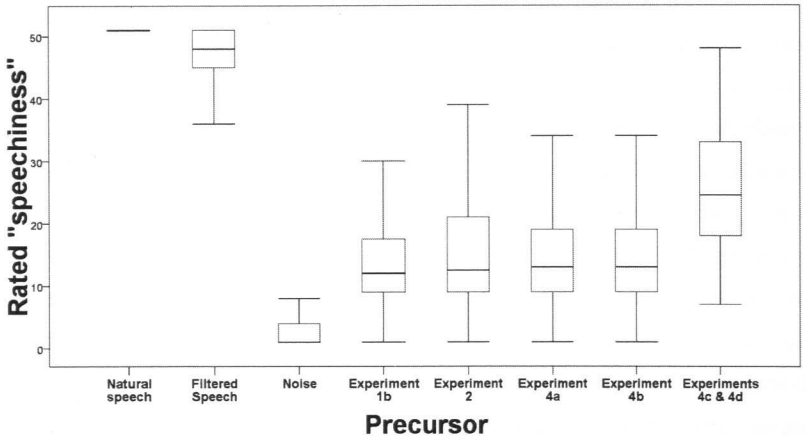

Figure 10. Experiment 5: Box plots representing "speechiness" ratings on a 51-step scale (0 = non-speech; 51 = speech) for natural speech, band-pass filtered speech, noise, and the precursors used in Experiments 1b, 2, 4a, 4b and 4c & 4d.

Furthermore, the precursors of the experiment with all but spectral rotation (Experiments 4c and 4d) were rated as far more speech-like than the precursors of Experiment 1b ($b = 12.44$, $p < 0.001$). This was found despite the fact that the normalization effects in Experiment 1b and 4d are numerically similar, with a smaller b value for the context effect in Experiment 4d than in Experiment 1b. These results provide further evidence against an account which suggests that normalization depends on the perceived speechiness of the precursors.

General Discussion

It was tested whether listeners take a precursor signal into account when categorizing speech or non-speech targets. Normalization was found, that is, an influence of a precursor on the perception of a target, with both speech and non-speech sounds. For both speech and non-speech targets, normalization varied with the exact acoustic shape of the precursor however. If the precursor did not contain

speech-like prosody, that is, amplitude and F0 variations, normalization was severely reduced for speech sounds (Experiment 4c and 4d compared to Experiment 1a) and normalization disappeared for non-speech sounds (Experiment 2 compared to Experiment 1b). A final experiment tested whether the amount of normalization varied with the perceived "speechiness" of the precursors. This was not the case. Spectrally-rotated stimuli with prosodic variation but an atypical spectral tilt created similar normalization effects to stimuli with a speech-like spectral tilt but without prosodic variation. Nevertheless, the former were rated as much less speechlike than the latter.

One point of departure for the current study was the finding by Watkins (1991) and Watkins and Makin (1994) that normalization at central processing levels is restricted to stimuli containing spectrotemporal variation. Our results reveal additional restrictions, because only two out of the five manipulated stimulus sets induced normalization (see Table 1). One of these was a spectrally rotated version of the otherwise intact speech stimuli. These signals therefore contained speech-like acoustic aspects (prosodic aspects, see Table 1 for details). This result also shows that it is not necessary for precursor signals to consist of natural or interpretable speech for normalization to occur. Once stimuli contained speech-like acoustic aspects, central normalization processes did take place, probably in the form of compensation for LTAS. It thus seems that acoustic similarity to speech is a prerequisite for the LTAS of the precursors to influence the perception of subsequent targets. Interestingly, this prerequisite also applies if the materials contained manipulated speech materials, as in Experiments 4c and 4d. These precursor signals contained identifiable steady-state vowels but lacked prosodic variation.

To account for the current set of results we suggest that, in compensation for LTAS at central processing levels, learning plays an important role. When listeners are confronted with different speakers in daily life, they might learn that the acoustic properties of speakers remain relatively stable. Given extensive exposure, listeners could therefore learn that taking the LTAS into account is beneficial while listening to speech. A striking parallel can be drawn with findings by Johnson, Strand and D'imperio, (Johnson, Strand, & D'Imperio, 1999) who report an auditory identification shift for a *hood* - *hud* continuum for listeners who merely imagined listening to a male or female speaker. In this situation normalization is thus even possible without any acoustic precursor input. A learning approach can account for

these findings as it suggests that, through experience, listeners acquire the ability to normalize auditory input for speaker characteristics, be it LTAS or learnt gender characteristics. A learning account can also accommodate the finding that the amount of central normalization varies with the exact nature of the precursor. Apparently, precursors with little prosodic variation are judged as less relevant for upcoming information and, therefore, their influence on the perception of the target signal is diminished or even obliterated.

An alternative to a learning account might be an enhanced auditory account in which the auditory properties directly determine (i.e., without learning) whether compensation for LTAS would take place. Listeners are sensitive to acoustic change rather than to constancy (Kluender, Coady, & Kiefte, 2003; Kluender & Kiefte, 2006). Perceptual properties of static information are reduced in order to increase sensitivity to more informative information. This explains several findings of contextual influences such as those by Watkins and Makin (1994, 1996) and Kiefte and Kluender (2008). From this viewpoint it could be that the lack of normalization effects that was observed by Watkins (1991) with filtered noise precursors is the result of a reduction of the perceptual effect of the precursors as a result of their constant nature. However, the materials in Experiments 4d and 2, for example, did contain considerable spectrotemporal variation, such as the movement of F_1 and F_2 (or its non-speech counterpart in the non-speech experiments), but did not give rise to a normalization effect (Experiment 2), or only to a strongly reduced one (Experiment 4d). There were also considerable spectral differences between the precursors and the targets. Furthermore, the fact that listeners reliably categorized the stimuli in all experiments shows that our precursor sets did not lead to a general lack of sensitivity to the F_1 properties of our stimuli (or the properties of the spectrally rotated counterpart of F_1).

The current set of experiments thus show that an alternative proposal, not dependent on learning, would have to go beyond the prerequisite of spectrotemporal variation that was proposed by Watkins (1991) and Watkins and Makin (1996). Such a proposal would have to account for the lack of normalization effects found in Experiments 2, 4a and 4b, and the reduction of the effect in Experiment 4d compared to Experiment 1a. Other potential explanations of why normalization effects were not observed here might focus on fine-grained spectro-temporal properties and/or spectral tilt of the precursors. For instance, the rising spectral tilt in some of our non-speech materials is unlike most sounds in nature and might induce perceptual effects unlike

those found with natural sounds. In analogy to the approach that was taken in Experiment 4, such factors could be investigated by analyzing the effect of individual acoustic aspects of the signal. At this point, however, an obvious single acoustic candidate property fails to emerge from the data.

Although more research is needed on a learning-based account, it nevertheless seems to be the simplest way to extend a general auditory theory so as to also explain the current findings. Importantly, however, we do not want to suggest that all perceptual normalization is a consequence of learning. Instead, the total amount of normalization that is found in speech perception is likely to be a combination of normalization/compensation processes that take place at different stages of processing (Holt & Lotto, 2002; Mitterer, 2011; Watkins, 1991). We attempted to focus on normalization that takes place at higher and more central levels in the processing stream by introducing a 500 ms precursor-target interval. The importance of this manipulation becomes clear from findings of normalization effects obtained with speech-shaped signal-correlated noise sounds (Watkins, 1991; Watkins & Makin, 1994). Watkins (1991) found a large categorization shift at 0 ms precursor-target interval when such precursor sounds were presented ipsilaterally, but a complete absence of compensation when the precursor signals were presented contralateral to the target sounds (indicating that compensation took place at a peripheral level of processing). When these precursors were presented binaurally with a precursor-target interval of 160 ms, there was still a shift of about half the size of the shift that was found with ipsilateral presentation. As Watkins (1991) suggests, it is thus likely that at 160 ms there is still some compensation due to peripheral mechanisms. This peripheral effect might be of the same type as that which causes the auditory after-images found by Wilson (1970). Central compensation mechanisms add to the normalization that is the result of those earlier, more peripheral mechanisms.

A possible intermediate stage is formed by perceptual contrast effects (Holt & Lotto, 2002). These effects (also described as compensation for coarticulation effects) are compensation that also give rise to a similar categorization shift as is found with extrinsic normalization. These effects have usually been investigated by categorization of consonant pairs such as /ga/ vs. /da/, and show that listeners more often interpret an ambiguous sound from a /ga-da/ continuum as /ga/ when presented after /al/ than when presented after /ar/. Like extrinsic normalization, this shift can be explained as arising from contrastive spectral characteristics in the preceding syllable.

A preceding syllable with more energy in higher frequency regions leads to the perception the low-frequency target and vice versa for a preceding syllable with more energy in lower frequency regions. Holt and Lotto (2002) and Lotto, Sullivan and Holt (2003) investigated the time-course over which local contrast effects generally decay and found that they were restricted to durations no longer than 400 ms however, which makes it unlikely that these effects had a large contribution in the effects reported in this manuscript. Importantly, however, Holt and Lotto (2002) showed that these effects are also obtained with contralateral presentation. They thus argued that these effects are different from peripheral adaptation effects.

These local contrast effects do appear to operate in a general auditory way. Lotto, Kluender and Holt (1997) have reported demonstrations of local contrast effects in birds. Furthermore, Holt (2006b) found compensation effects with notched-noise precursor stimuli (although peripheral effects were not controlled for as stimuli were presented bilaterally and with a 50 ms interval). Either of these two relatively early levels could also be the level of operation for the mechanism causing effects of speech and musical precursors on immediately following musical targets (Stilp, et al., 2010). Interestingly, Stilp et al. (2010) also address the issue of learning with these materials. They report that participants' musical experience did not influence the size of the compensation effect that was found. These combined findings suggest that effects at these early levels of processing are indeed of a general nature, indifferent to the exact acoustic nature of the preceding signal, and not dependent on learning.

It appears, however, that these two relatively early compensation mechanisms do not influence categorization with precursor-target intervals of 500 ms or more. If extrinsic normalization were due to an early compensation mechanism, that is indifferent to the fine-grained spectral and temporal properties of the signal but can apply over at least 500 ms, then normalization ought to have been found in Experiments 2, 4a and 4b (or, for that matter, with the signal-correlated noise materials used in Watkins, 1991). The materials in all these cases had the LTAS properties to induce compensation, yet no effect was found.

The current results, however, also demonstrate that the compensation effects of a possible learning mechanism are not restricted to speech per se. Acoustic signals that are, at a gross level, sufficiently similar to speech signals are subject to similar types of normalization. The fact that gross acoustic similarity is enough to induce normalization effects is in accordance with normalization effects that have been found

with temporally reversed precursor syllables (Watkins, 1991; Watkins & Makin, 1994) rather than normal sentences. An important additional implication of a learning approach is that learning to adjust perception for acoustic context signals is indeed by no means restricted to the processing of speech sounds. Any sound structure that shows LTAS constancy over time (and for which it would be beneficial to normalize, e.g., due to overlapping sound categories) could in principle evoke learned normalization processes.

In sum, we suggest that normalization for context at short ISIs is driven by automatic auditory processing that is independent of learning, while normalization at longer ISI is influenced by learning. This suggestion seems to be at odds with one particular set of findings, however. Holt (2005) reports an influence of a long sequence of steady sine wave tones on categorization of a subsequent ga/da continuum, despite an interstimulus interval of up to 1.3 seconds. There are two possible ways to interpret this discrepancy. A first explanation could be that the context effects found by Holt (2005) could be of a different type: Such effects have generally been investigated with changes in categorization of (transient) consonants, whereas our investigation focuses on shifts in identification for (more stationary) vowels. This is in line with a suggestion by Mitterer (2006b). Mitterer investigated compensation for coarticulatory lip-rounding in fricative perception, and found a pattern of results that was completely opposite to the pattern found for stops. Just as in the current case, this suggests that normalization for transient stimuli (stops) may be different from normalization for stationary stimuli (vowels and fricatives). Alternatively, Holt (2005) argues that the persistence of the effects obtained with these acoustic histories, compared with the rapidly diminishing effects obtained with temporally adjacent contexts (Holt & Lotto, 2002; Lotto, et al., 2003), supports the idea that contrast effects may exist at multiple time scales. This supports the idea that local contrast effects and normalization effects over longer time scales are due to functionally different processing levels. This makes it possible that Holt (2005) reports on effects that should be attributed to the same processing level as the effects reported here. The acoustic context in the study of Holt (2005) contained an alteration of tones and silences, which may be sufficient prosodic variation for learned normalization to be engaged.

Another apparent contradiction with our suggested role of familiarity stems from a finding by Remez, Rubin, Nygaard and Howell (1987), who reported extrinsic normalization effects (also at an ISI of 500 ms) with sine wave replicas of precursor sentences and targets similar to those used by Ladefoged and Broadbent (1957). On the one hand, such sine wave models lack the acoustic complexity of speech signals, and it is unlikely that listeners will have been exposed to these kinds of materials very often. On the other hand, sine wave models can be interpreted as speech, and therefore appear to share some crucial auditory and prosodic aspects with speech as well. In addition to the availability of formant-like structures, sine wave models have specific time-varying phonetic characteristics such as low amplitude parts and syllable amplitude onsets and offsets (attack and decay structures) that are typical of speech sounds. As such, sinewave replicas of speech sounds could be sufficiently similar to familiar speech sounds to engage a learned normalization process.

To summarize, the current set of results suggests that extrinsic normalization at central processing levels is not exclusively the result of an all-purpose auditory compensation mechanism that is indifferent to the exact nature of those stimuli. The non-speech precursors that were spectrally complex but which had relatively little similarity to speech did not produce normalization effects on their non-speech targets. Importantly, however, normalization was not restricted entirely to speech stimuli: Once non-speech stimuli were substantially similar to speech materials on an acoustic level, normalization effects were found. We suggest that the simplest and most parsimonious way to explain these results is to assume that an important component of the perceptual normalization of vowels over longer time-spans could therefore be an auditory mechanism that has been acquired over a lifetime of experience with different speakers or acoustic events. This perceptual mechanism could be a learned response to the covariations in natural sound patterns. Learning is a useful way to deal with the contextual influences that have not been resolved at lower levels in the processing stream. Lower-level compensation mechanisms deal with covariations at shorter latencies, and can do so because, at short latencies, it is more likely that a new sound originates from the same source. At longer latencies, however, stability of a sound source is less likely and therefore general mechanisms can be potentially harmful if they compensate perception of a new sound using the wrong source characteristics. A learning-based mechanism is a relatively simple way to deal with more situation-dependent covariations at longer latencies because such a mechanism

can adjust to specific situational demands. Longer-term extrinsic normalization appears to operate as one of an array of other contextual mechanisms such as lower-level compensation mechanisms (Wilson, 1970), intrinsic normalization (Nearey, 1989), audio-visual integration (Massaro & Jesse, 2007) and lexically-guided retuning of speech perception (Norris, et al., 2003). These mechanisms act in concert to resolve variability in speech signals.

Appendix A

Pretest

This pretest was designed to establish the step size needed in order for listeners to reliably distinguish the /pɪt/ and /pɛt/ stimuli, and the non-speech versions of these stimuli, using a staircase procedure.

Method.

Eight Dutch participants from the Max Planck Institute participant pool were recruited. They received a monetary reward for their participation. Sounds were created ranging from [pɛt] to [pɪt] (using the base sound and construction method described under Experiment 1). These sounds were combined in pairs. The first level consisted of [pɛt] (0 Hz F_1 decrease) and [pɪt] (200 Hz F_1 decrease, relative to [pɛt]). This level will now be referred to as the "0-200" pair. The other levels consisted of increasingly smaller differences (e.g., "10-190", "20-180"). The difference decreased in steps of 20 Hz until it was 100 Hz. It then decreased in steps of 10 Hz until it was 50 Hz. Then it decreased in steps of 4 Hz until it was 22 Hz, and finally it decreased in steps of 2 Hz until it was 0 Hz. There were 28 steps in total. For catch trials, one of the sounds from each level appeared twice (e.g., 87-87). To create the non-speech stimuli, all of these materials were spectrally rotated.

Participants were instructed that the experiment was designed to establish the limits of their hearing. They were told that, on every trial, they would hear two sounds which were either the same or different, and they had to respond by pressing response keys labeled "Hetzelfde" (same) or "Verschillend" (different). Participants judged both the speech version and the spectrally-rotated version (order was randomized over participants). Each version started with a frequency difference of 200 Hz (the "0-200" pair).

When participants correctly responded to a block of stimuli (four pairs; two same, two different, in randomized order, from the same difficulty level) they would move to the next, more difficult, level, until they reached their threshold (an upward run). If a participant responded incorrectly, the difficulty level would decrease by two levels, every time until the participant completed a block correctly (a downward run). The increase in difficulty in an upward run involved large level changes before the first and second downward runs (5 and 3 levels, respectively), and increases of only one level thereafter (up to a total of 7 upward runs). A participant's discrimination

level was calculated as the average level at which that participant started his or her upward runs, based on the last 5 upward runs.

Results.

The results are displayed in Figure 11, which displays the average lowest level at which an individual participant started an upward run. The difference levels to be used in the main experiments were selected to be those which all eight participants could discriminate (i.e., the lowest levels at which all participants showed, on average, discrimination). For the speech version this level was set at a difference of 60 Hz. For the spectrally rotated version this level was set at a difference of 200 Hz.

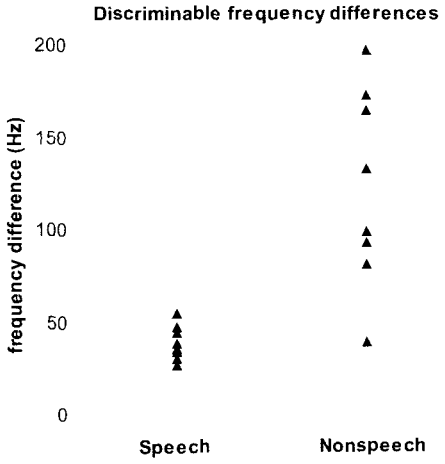

Figure 11. Pretest: Participants made same-different judgments on pairs of stimuli in a staircase procedure. Triangles represent individual participants and show the average level at which that individual reliably discriminated the stimuli

Appendix B

Table: Number of Blocks Needed to Reach Criterion in the Training Phases of Each Experiment

	Training phase						
	Part 1		Part 2		Part 3		
Experiment	Mean	Max	Mean	Max	Mean	Max	Participants
1a	9	29	4.5	11	3.1	4	8/9
1b	6.8	16	4.6	9	6.4	20	8/10
2	6.1	25	3.4	6	3.3	4	8/8
4a	4.3	7	3.4	6	5.8	17	8/11
4b	4.5	12	3.3	4	6.9	18	8/8
4c	3.3	4	3	3	3	3	8/8
4d	6.5	12	4.3	10	5	13	8/10

Note. Mean number of blocks (mean) and maximum (max) number of blocks across participants are shown for each part of the training phase. Also shown are the number of participants that reached criterion, out of the total number of participants that were tested.

Normalization for vocal tract characteristics does not depend on attention

---

Chapter 3

Sjerps, M. J., McQueen, J. M., and Mitterer, H. (in preparation). Normalization for vocal tract characteristics does not depend on attention.

Abstract

This study investigated the influence of attention on extrinsic vowel normalization. Vowel normalization has been measured as shifts in categorization of stimuli on an F_1 vowel continuum, presented after precursors with a high versus a low F_1. Such shifts have previously been found with spectrally-rotated speech and with speech that was otherwise manipulated to make it sound less speech-like. These earlier experiments were replicated here, except that listeners had an additional task focusing attention on the precursors. Compensation effects were not stronger when listeners paid attention to the precursors, suggesting that vowel normalization is mainly determined by bottom-up signal characteristics.

Introduction

When listening to speech, listeners compensate for the vocal tract characteristics of the speaker in a preceding sentence (Ladefoged & Broadbent, 1957). When listeners categorize targets that lie on an F_1 continuum (such as /pɪt/ - /pɛt/) which are preceded by precursors that are manipulated to have either a high or a low F_1 contour, normalization effects are observed as a shift in categorization of the targets. This compensation mechanism has been argued to have a mainly auditory basis (Watkins & Makin, 1994, 1996). For instance, when categorizing target vowels spoken by a male speaker, listeners compensate for the vocal tract of a precursor even when this was spoken by a female (Watkins, 1991). Moreover, listeners compensate in a similar way to speech as to spectrally-rotated speech, even though they generally interpret spectrally-rotated speech as not being speech (Chapter 2 [Sjerps, Mitterer, & McQueen, 2011]) (spectral rotation changes the frequencies of the formants but preserves the spectrotemporal complexity of the signal Blesser, 1972). Similarly, normalization effects with tone sequences as precursors have been reported (Holt, 2005).

Recent investigations of normalization effects with non-speech materials, however, have questioned the purely auditory nature of normalization effects (Chapter 2). A number of manipulations were applied to speech stimuli that made the new stimuli unlike speech. These non-speech manipulations were always applied to the high- F_1 and the low- F_1 precursor. One of these manipulations was spectral rotation. Spectrally rotating a precursor causes the F_1 contour to be moved to a completely different frequency region. To keep the acoustic relation between the precursors and the targets similar, the targets were also spectrally rotated. The question was whether these manipulations would influence the normalization effects. It was found that some of the non-speech precursors that were created did not induce normalization effects. For instance, by removing low amplitude parts (such as silent closures in stops), by setting the pitch contour to a fixed value (at ~224 Hz), by temporally reversing the syllables and setting them to an equal amplitude, and by spectrally rotating them around 1250 Hz, a non-speech precursor signal was created that did not induce normalization.

The findings reported in Chapter 2 therefore show that normalization processes do not always occur. They suggest that the occurrence of normalization depends on some specific auditory properties of the precursor (apart from the Long

Term Average Spectrum, or LTAS, relation between precursor and target). Alternatively, however, the discrepancy between normalization with speech signals versus no normalization with non-speech signals could reflect an attentional effect. In particular, it could reflect an influence of how relevant listeners judge the precursor signal to be in relation to the perception of the following target. Based on perceived relevance, listeners may pay more or less attention to the precursors. Because speech signals are naturally more informative than non-speech signals, it is not unlikely that listeners will pay more attention to speech than to non-speech stimuli. Moreover, there might be other properties of non-speech precursors that decrease the amount of attention that participants pay to them. Lack of normalization with non-speech signals may thus reflect listeners' lack of attention to those signals.

The current experiments were therefore set up to test the hypothesis that the occurrence or amount of normalization is influenced by whether listeners pay overt attention to the precursors. This allowed us to investigate whether the amount of normalization that is observed is determined mainly by bottom-up signal properties or also by attentional influences.

To investigate the contribution of attention to extrinsic vowel normalization, two experiments were run. These experiments were replications of two experiments reported in Chapter 2 but participants had an additional task that focused their attention on the precursor. Although there was no normalization in some experiments with non-speech signals in Chapter 2, two types of acoustic manipulation *did* produce normalization effects. In the first of these experiments, precursors and targets were both spectrally rotated (Experiment 1b in Chapter 2) These signals did not sound like speech, but normalization was found. In the second experiment (Experiment 4c in Chapter 2), speech targets were preceded by a precursor that was manipulated by removing its low amplitude parts, setting its pitch contour to a fixed value, temporally reversing the syllables and setting them to an equal amplitude (i.e., in the same ways as the most extreme manipulation described above, *except* that there was no spectral rotation). For the targets the pitch was also set at a fixed value in order to create similarity between the precursor and target. Although the precursor signals were manipulated extensively (and although this led to a non-speech percept, especially for the spectrally-rotated materials), normalization was found.

The two non-speech precursor-target combinations that did elicit compensation effects in Chapter 2 were used here, rather than those that failed to

show effects. This choice was made because signals that have been shown to induce small effects will probably be more susceptible to an attentional manipulation than those previously showing null effects.

The procedure was very similar to that in the experiments reported in Chapter 2. Participants were asked to categorize spectrally-rotated speech targets (Experiment 1) or speech targets that had a flat pitch (Experiment 2). These targets were preceded by precursors. Within an experiment, a precursor could have either a high F_1 or a low F_1 (or spectrally-rotated analogs of these formant values in Experiment 1). Additionally, the precursors in Experiment 2 were manipulated in a number of ways (no low amplitude parts, flat pitch contour, reversed syllables of equal amplitude; see below for further details). Critically, and in contrast to Chapter 2, an additional task encouraged participants to attend to the precursors. In this additional task, participants were asked to refrain from responding to the target whenever the precursor had a dip in amplitude (these catch-trial precursors were presented occasionally throughout the experiment).

In summary, the main goal of this study was to test the influence of an attentional task on the size of vowel normalization effects. Additionally, these experiments tested whether the normalization effects found with manipulated signals in Chapter 2 were robust enough to be replicated.

Experiments

Participants

Twelve participants were recruited for Experiment 1 and 8 participants were recruited for Experiment 2. Experiment 1 required 4 additional participants because there was considerable individual variation in effect size for the first 8, although the average effect was of similar magnitude and in the same direction as the average after 12.

Stimuli

Experiment 1.

Base target sounds consisted of a six step [pɪt] to [pɛt] continuum (an F_1 distinction). The steps were created by lowering the F_1 from a recorded instance of /ɛ/ in 6 steps of 40 Hz. The targets were then all spectrally rotated around 1250 Hz. Spectral rotation is a transformation that rotates the spectral make-up of a complex signal around a central frequency, such that the information in the high frequency ranges trades places with the information in the low frequency ranges. The precursors

were based on a Dutch sentence (*"Op dat boek staat niet de naam"*, lit. on that book is not the name). This sentence was manipulated to have either a low F_1 (-200 Hz) or a high F_1 (+200 Hz). These versions of the precursor were then spectrally rotated in the same way as the targets.

Experiment 2.

Targets materials were the same as in Experiment 1, with the exception that they were *not* spectrally rotated and that they had a flat pitch level to increase similarity between the precursors and targets. The pitch was flattened using the overlap-add method for resynthesis in Praat (Boersma & Weenink, 2005). The base speech precursors (i.e., the versions prior to spectral rotation) from Experiment 1 were used for Experiment 2. Several additional manipulations were applied to both the low-F_1 and the high- F_1 precursors. The signals were modified to have a flat pitch at the average value of the speech materials (223.8 Hz) using the same method as was used for the targets. Each of these signals was divided in high and low amplitude parts (see Figure 1). All the high amplitude parts were temporally reversed (e.g., the first digital sample of a part became the last sample of the new "reversed part" and vice versa) and equalized in amplitude relative to each other. All low amplitude parts were excised and discarded.

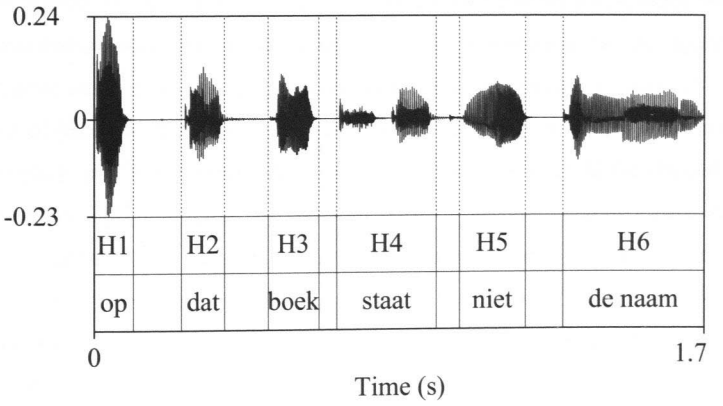

Figure 1: The annotated base sentence, divided into high (H) and low amplitude parts.

Procedure

The training and testing procedures were identical in both experiments and were also identical to those used in Chapter 2 except for the addition of the catch trials.

Training.

As the participants for Experiment 1 were presented with novel non-speech stimuli, they first had to undergo a three-phase training protocol to familiarize them with these materials. The same procedure was applied for Experiment 2 to keep the amount of exposure across the experiments similar. In each training phase, participants had to reach a performance criterion to go on to the next training phase or, after the third training phase, to the testing part of the experiment. During all three training phases, but not at test, visual feedback ("correct" (correct) or "fout" (incorrect)) appeared on a computer screen after each trial. The first training phase consisted of a discrimination task using only the endpoint targets (a same-different task; criterion: for three consecutive blocks, seven out of eight correct). The second phase consisted of a categorization task with the endpoint targets (with the options "A" and "B"; criterion: for three consecutive blocks, nine out of ten correct). Listeners were told that they had to find out which target belonged to which button ("A" or "B"). The third training phase consisted of a categorization task that was similar to the second phase, with the addition that the targets were preceded by a neutral version of the precursor (i.e., a version with no F_1 manipulation). Additionally, however, during the third training phase there were catch trials that indicated to participants that they should refrain from responding. These catch trials were not included in Chapter 2. Catch trials could be recognized by a two-syllable long dip in amplitude of 20 dB. Catch trials (pseudo) randomly varied in where the amplitude dip would occur (see Figure 1, H2&H3, H3&H4, H4&H5 or H5&H6). In order not to change the criterion relative to Chapter 2, erroneous button presses on the catch trials did not influence whether participants could pass to the test phase. One catch trial was presented every 10 trials.

Testing.

In both experiments, the six target steps were each played after both the high and low precursors (in random order) for 15 repetitions, resulting in 180 test trials (with two self-paced pauses). Trials were presented without feedback. Participants categorized the targets with the same two buttons as those used during the second and

third training phases ("A" and "B"). In addition to the test trials, the testing phase also contained catch trials that were constructed in the same way as those in the last training phase. On every block of twelve trials (6 steps x 2 precursors) two additional catch trials (one with a high F_1 and one with a low F_1) were presented. After such precursors, the middlemost step of the continuum was presented (halfway between steps 3 and 4). As in the training phase, participants had to refrain from responding on catch trials. All stimuli were presented with a 500 ms silent interval between a precursor and the following target.

Results

The data were analyzed using linear mixed-effects models in R (version 2.6.2, R development core team, 2008, with the lmer function from the lme4 package of Bates & Sarkar, 2007). Different models were tested in a backward elimination procedure, starting from a complete model. All factors were numerical and centered around 0. These included the factors Step (levels: -2.5 to 2.5 in steps of 1), Precursor (levels: low F_1 = -1 vs. high F_1 = 1), Block (15 stimulus repetitions: levels -7 to 7 in steps of 1) and their possible interactions. A comparison with the effects reported in Chapter 2 was made by including the data from the earlier study and the factor Attention in the model (levels: data from Chapter 2 = -1 vs. current results = 1). Non-significant predictors were taken out of each analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was found by means of a likelihood ratio test.

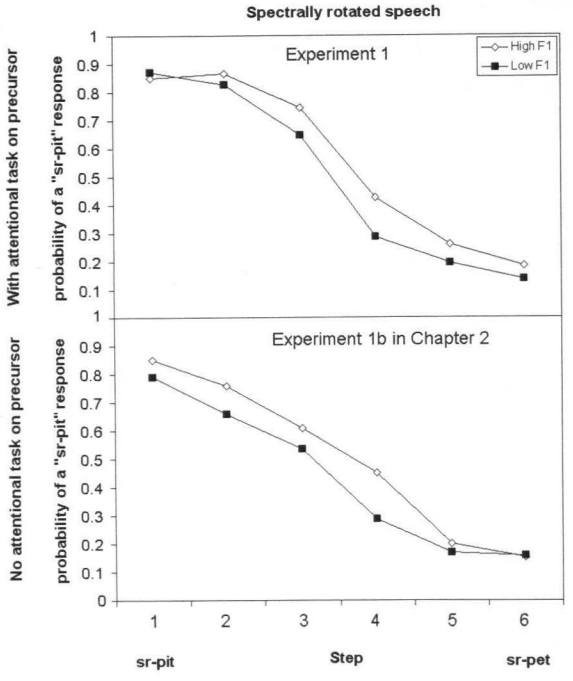

Figure 2: Probability of spectrally rotated /pɪt/ ("sr-pit") responses to stimuli on a sr-/pɪt/ to sr-/pɛt/ continuum. Targets were preceded by precursors that were manipulated to have a high or a low F_1 and that were also spectrally rotated. Top panel: data for results reported here, in a task where participants were encouraged to pay attention to the precursors. Bottom panel: data for results reported in Chapter 2 with no attentional task.

Experiment 1.

On seventy-two percent of the catch trials participants correctly refrained from responding. Only the data for the non-catch trials were further analyzed. The top panel of Figure 2 displays the results obtained here, with a compensatory normalization effect: more high F_1 responses after a low F_1 precursor. This pattern was confirmed in the statistical analysis. The best-fitting model included main effects for the factors Step ($b = -0.844$, $p < 0.001$) and Precursor ($b = 0.365$, $p = 0.001$). The latter reflects a normalization effect.

The lower panel of Figure 2 displays the results from the corresponding experiment in Chapter 2 (reported as Experiment 1b). They revealed effects in the same compensatory direction. In the comparison of the effects obtained here and in Chapter 2, modeling settled on main effects for the Intercept ($b = -0.206$, $p = 0.025$)

and the factors Step ($b = -0.785$, $p < 0.001$) and Precursor ($b = 0.377$, $p < 0.001$). An interaction was found between the factors Step and Attention ($b = -0.126$, $p = 0.023$) which reflects the fact that participants responded more categorically (i.e., they showed a steeper categorization curve) in this experiment than in Experiment 1b reported in Chapter 2. There was no interaction of Precursor and Attention, suggesting the addition of the attentional task had no effect on normalization.

Experiment 2.

On sixty-two percent of the catch trials participants correctly refrained from responding. Analyses were carried out on the non-catch trial data. The top panel of Figure 3 displays the results obtained here, with no visible compensation effect.

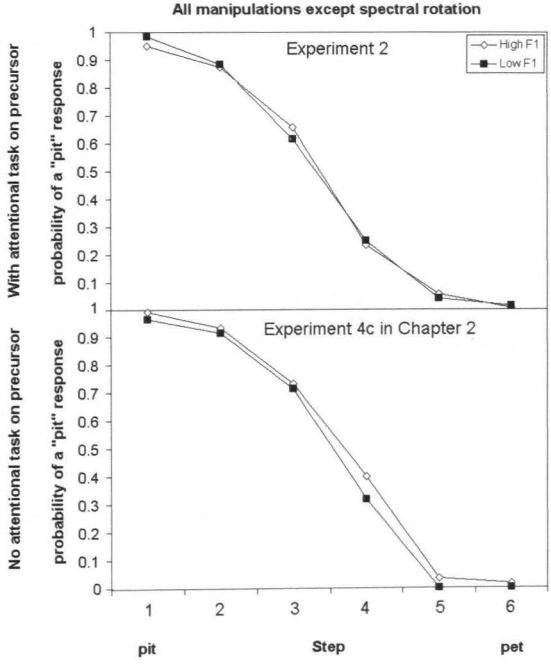

Figure 3: *Probability of /pɪt/ ("pit") responses to stimuli on a /pɪt/ to /pɛt/ continuum. Targets were preceded by precursors that were manipulated to have a high or a low F_1. They were further manipulated to have a flat pitch, no low-amplitude parts and temporally reversed syllables that had equal amplitudes. Top panel: data for results reported here, in a task where participants were encouraged to pay attention to the precursors. Bottom panel: data for results reported in Chapter 2 with no attentional task.*

The absence of an effect was confirmed in the statistical analysis. The final model consisted of the single main effect of Step ($b = -1.690$, $p < 0.001$). When the factor Precursor was added to this model, it did not produce an effect ($b = -0.010$, $p = 0.953$). This shows that no normalization effect was observed for this experiment.

The bottom panel of Figure 3 displays the results from Experiment 4c in Chapter 2, where a small effect in the compensatory direction was found. When the current experiment was compared to the previous experiment, the optimal model consisted of a main effect of Step ($b = -1.716$, $p < 0.001$) and an interaction between Step and Attention ($b = 0.924$, $p < 0.001$), which reflects the fact that listeners responded more categorically in (Chapter 2). The model also consisted of three-way interactions between Step, Precursor and Block ($b = 0.060$, $p = 0.038$), which is a reflection of an effect found in Chapter 2, and between Step, Block and Attention ($b = -0.077$, $p = 0.009$), which reflects the fact that participants in Experiment 2 became more categorical towards the end of the experiment than those in Experiment 4c in Chapter 2. No overall effect of Precursor was found, nor an interaction of Precursor with Attention (although there were trends in the direction of an overall effect of Precursor and an interaction of Precursor with Attention such that the effect reported here was smaller).

Discussion

This study investigated the role of attention in extrinsic vowel normalization. Two experiments reported in Chapter 2 were replicated here with an additional attentional task. Listeners categorized speech targets on a [pɪt] to [pɛt] continuum (Experiment 2) or spectrally rotated versions of these (Experiment 1). These targets were preceded by precursors, in two conditions, manipulated to have an F_1 contour that was increased or decreased by 200 Hz. In Experiment 1 these precursors were also spectrally rotated. In Experiment 2 the precursors were not spectrally rotated but manipulated in a number of other ways (no low amplitude parts, flat pitch contour, reversed syllables of equal amplitude). Furthermore, in contrast to Chapter 2, listeners in both experiments were encouraged to pay attention to the precursors through the inclusion of an additional task: They had to refrain from responding when a dip in amplitude occurred in the precursor.

The attentional manipulation did not increase the amount of normalization. A normalization effect was found in Experiment 1 and this effect was of a similar size to the effect in Chapter 2. Although no significant normalization effect was found in

Experiment 2, the difference between this experiment and that in Chapter 2 was not significant. It appears that paying attention to a precursor sound does not change the precursor's influence on categorization of subsequent sounds (and Experiment 2 suggests that, if anything, attention might make the effect smaller). The second important finding of this study is that normalization with spectrally-rotated speech was replicated (Experiment 1). This confirms that robust normalization effects can be observed with non-speech signals (and across 500 ms precursor-target intervals).

These two conclusions, however, give rise to an apparent contradiction. It has been found that speech signals that have been manipulated extensively (to make them very unlike speech) do not give rise to normalization effects, while precursors that have undergone less extensive manipulations do produce effects (Chapter 2). There is thus an influence of the speech/non-speech dimension on the size or occurrence of the normalization effect. It seems likely that a major difference between speech and non-speech is the way in which we attend to them at a conscious level. But if the effect of degree of speech-likeness reflects this attentional difference, why did the present attentional manipulation have no effect? We suggest that the resolution of this apparent contradiction reflects learning rather than attention. We propose that, through experience, listeners have acquired knowledge about the low-level characteristics of speech signals. This includes the knowledge that it is beneficial to compensate for the acoustic properties of the source. When listening to a given speaker, for example, such a process is highly beneficial because overall vocal tract properties of speakers remain relatively stable. For other signals there may be less acoustic stability. Perceptual learning at very low levels of processing has been reported in the domain of speech pitch perception (Krishnan, Xu, Gandour, & Cariani, 2005). We suggest that, at low levels of processing, the perceptual system has gained experience with the spectrotemporal characteristics of speech and has learnt to process subsequent signals with similar spectrotemporal complexity in a non-independent fashion. This can partly be done by taking the spectral characteristics of preceding contexts into account (Kluender & Kiefte, 2006). This means that although signals are required to be similar to speech for normalization processes to occur, the similarity to speech is determined based on bottom-up signal characteristics and not on attentional responses to the signal.

To conclude, the current study provides a replication of normalization effects with non-speech signals that were created by spectrally rotating speech. Moreover, it was found that attention did not influence the amount of normalization that is observed. This shows that attentional processes in speech perception play little to no role in extrinsic vowel normalization. The size of normalization effects and indeed their very occurrence appear to be determined mainly by bottom-up signal properties.

Compensation for vocal tract characteristics across native and non-native languages

---

Chapter 4

Sjerps, M. J., and Smiljanic, R. (submitted). Compensation for vocal tract characteristics across native and non-native languages.

Abstract

Perceptual compensation for speaker vocal tract properties was investigated in four groups of listeners: native speakers of English and native speakers of Dutch, native speakers of Spanish with low proficiency in English, and Spanish-English bilinguals. Listeners categorized targets on a [sofo] to [sufu] continuum. Targets were preceded by sentences that were manipulated to have either a high or a low F_1 contour. All listeners performed the categorization task for precursors in Spanish, English and Dutch. Results show that listeners from each of the four language backgrounds compensate for speaker vocal tract properties. Listeners also compensate when they listen to stimuli in another language. The amount of compensation that was observed, however, differed for specific combinations of the speakers' background language and the language of the stimuli that they were listening to. These results show that patterns of compensation are not fully determined by auditory properties of precursor sentences.

Introduction

Across-speaker variation in speech production due to vocal tract differences has been well documented (Peterson & Barney, 1952). These differences contribute to variation in the exact acoustic-phonetic properties of speech sounds. For instance, a speaker with a long vocal tract will have a lower first formant (F_1) when uttering the vowel /o/ relative to a speaker with a short vocal tract. Variation due to differences in the size of speakers' vocal tracts can be problematic for listeners, because formant frequencies are an important determinant of vowel identity. In principle, this would suggest that a listener would be inclined to misperceive an intended /o/ for /u/ (that has a lower F_1 in general) when listening to a speaker with a longer vocal tract.

Listeners, however, can use additional information in the speech signal to overcome potential misperceptions. Such consists of, for example, both the higher formants and the speaker's pitch (Johnson, 2005; Nearey, 1989). This type of compensation has been termed intrinsic normalization as the necessary information is available within the targets vowel itself (Nearey, 1989). However, listeners can also utilize spectral information available in the context, that is, *outside* the target vowels, to guide their vowel categorization. Ladefoged & Broadbent (1957) showed that when listeners were categorizing sounds halfway between [ɛ] and [ɪ] (that have a high and a low F_1 respectively), they heard more sounds as /ɪ/ when the target word was preceded by a sentence with a high F_1 than with a low F_1. This showed that listeners compensated for the vocal tract properties of a speaker as revealed in a preceding sentence. It is unclear, however, whether such effects are largely caused by general auditory processes (and as such are mainly dependent on signal properties) or by a listener's compensation for phonetic properties of a specific speaker's vowel space. The latter suggests that these compensation effects are dependent on a listener's experience with auditory properties of phonetic categories in his or her native language. The current paper investigated whether the perceptual compensation with vowels is influenced by a listener's native language to address this question.

It has recently been suggested that the direction and amount of perceptual compensation is for an important part determined by the relation of the Long Term Average Spectrum (LTAS) between a precursor sentence and a target sound (Kiefte & Kluender, 2008; Kluender, et al., 2003; Watkins, 1991; Watkins & Makin, 1994, 1996). A precursor sentence with high amplitudes in high spectral regions will suppress the perceptual impact of those frequencies in a following target. When

context and target stimuli are presented with very short or no interstimulus intervals such effects have indeed been reported. These effect have been argued to partly arise in the peripheral system (Summerfield, Haggard, Foster, & Gray, 1984; Watkins, 1991; Wilson, 1970), and have, in analogy to processes in the visual system, been termed auditory after-images (Wilson, 1970). However, over longer precursor-target intervals and with contralateral stimulation similar compensation effects have been found (Holt & Lotto, 2002; Lotto, et al., 2003; Watkins, 1991) even though these manipulations reduce peripheral effects to a minimum or even abolish them (Summerfield, et al., 1984). This suggests that an important part of compensation effects have a central origin (Holt & Lotto, 2002; Lotto, et al., 2003; Watkins, 1991).

It has been argued, however, that these central processes still operate in a rather general auditory way. In speech perception, the logic is that when listeners categorize targets on, for example, an [ɪ] (that has a low F_1) to [ɛ] (that has a high F_1) continuum, a precursor with a relatively high F_1 will decrease the perceptual impact of higher frequencies in the subsequent targets. An ambiguous target sound that is preceded by a precursor with a high F_1 will then be perceived as one with a relatively low F_1 (and thus more /ɪ/-like) compared to the context in which the precursor sentence has a relatively low F_1. Perceptual compensation, then, acts in a fashion similar to applying a filter whose frequency response is the inverse of the LTAS of the precursor, to a target signal (Watkins & Makin, 1994, 1996). Mitterer (2006a) has shown that when targets are preceded by a precursor sentence with a manipulated F_2 contour (either a high F_2 or a low F_2), effects on categorization of the targets are found only for vowels that have their formants in the same region as the manipulated formant in the precursor (in this case in the mid-to-high front vowel region). This suggests that central compensation effects are largely determined by the spectral relation between a precursor and a target signal. This approach thus focuses on general signal properties that are largely independent of listeners background language, and even the speech status of stimuli (Stilp, et al., 2010).

This contrasts with the framework that has been termed "extrinsic vowel normalization" (Nearey, 1989). According to this approach, listeners estimate the formant values of a particular speaker's vowel space based on a preceding sentence. When hearing a subsequent speech sound from the same speaker, listeners perceive that target sound relative to a mental frame of reference. This approach focuses much more on listener's phoneme repertoire and therefore a listener's language background.

This finds some support in reports such as those by Watkins and Makin (1996), Watkins (1991) and Mitterer (2006a), who have shown that a signal correlated noise precursor induced significantly smaller compensation effects than a speech precursor that had the same LTAS, and sometimes even a lack of compensation effects with non-speech precursors. From a mental frame of reference standpoint, these findings suggest that only a small part of the compensation found with speech sounds is actually due to general auditory processes, while an important contribution is made by a language-specific process such as, possibly, a phonetic frame of reference. Watkins (1991) and Watkins & Makin (1996), however, have argued that the lack of compensation effects with noise precursors could be the result of the fact that the noise precursors did not contain spectrotemporal variation. The latter proposal, however, cannot account for the finding that, despite similar LTAS relations between precursors and targets, some non-speech precursor signals induce stronger compensation effects than others (Sjerps, Mitterer, & McQueen, 2011) even though they all had spectrotemporal variation. Notably, Sjerps et al, (2011) argued that those stimuli that were acoustically (although not necessarily perceptually) more similar to speech induced larger compensation effects. This suggests that language experience could have had an influence on compensation processes.

Perceptual compensation for speaker vocal tract properties has been shown in at least two languages, Dutch and English (Broadbent & Ladefoged, 1960; Ladefoged & Broadbent, 1957; Mitterer, 2006a; van Bergem, et al., 1988; Watkins, 1991; Watkins & Makin, 1994, 1996). Both of these languages have large vowel inventories, which result in a crowded vowel space with considerable overlap between instances of different vowel categories in the F_1/F_2 vowel space. For listeners of these languages it is beneficial to reduce this overlap by compensating for vocal tract characteristics. It remains unclear whether the same processes will apply in a language with a smaller vowel inventory such as Spanish. In Spanish, perceptual confusion among vowel categories is, presumably, less severe than the confusion observed in American English (Cutler, Weber, Smits, & Cooper, 2004). If the strength of compensation processes is for an important part based on language experience, one would expect that listeners of English and Dutch have learnt to compensate relatively stronger than listeners of Spanish. Furthermore, it has not been established to what extent compensation processes apply when listening to a second or an unfamiliar language. The current study investigated perceptual adaptation to vocal tract

properties in native speakers of Spanish, native speakers of American English and native speakers of Dutch when listening to target vowels embedded in sentences in all three languages. Testing the same participants with materials from different languages, while controlling for their experience with each language, allowed us to examine perceptual compensation patterns across different languages and across listeners' native and non-native languages. This design allowed us to investigate the contribution of language exposure on compensation processes.

In this study, the strength of compensation effects were tested by means of a target vowel pair that is shared among the three languages and that has no other vowels in between them in any of the stimulus languages. The pair that was considered most suitable was the /o/ - /u/ pair, a distinction that lies mainly on F_1 in the three languages. It should be noted, however, that this similarity is true in an abstract phonological sense, and only partly so in its phonetic realization. While F_1 and F_2 are in similar areas in vowel space, especially English has diphthongization for /o/ and /u/. For the current study, however, it is important that listeners from all three languages rely on F_1 for the distinction of /o/ and /u/ to some extent. All listeners heard target vowels on a [sufu] - [sofo] continuum, preceded by Spanish, English and Dutch precursors. These CVCV sequences were chosen because they are non-words in all three languages. Furthermore, the consonants /s/ and /f/ are produced similarly in all three languages. Critically, the precursors were manipulated to have a generally high or a generally low F_1. This design allowed us to examine four hypotheses.

The first hypothesis is that compensation processes apply to all speech sounds (and to the same extent), irrespective of the language in which the precursor and target are uttered. Such a proposal predicts that all listeners, Spanish, English and Dutch, normalize for all precursor sentences (thus irrespective of the stimulus language). The amount of compensation is then influenced only by the spectrotemporal characteristics of the signals. Note that the precursor signals for the different languages were necessarily different. The LTAS-based influence of the different materials will therefore inevitably differ to some extent. Our focus here is, therefore, on whether listeners from different language backgrounds are influenced by these precursors to a different extent.

The second hypothesis, in contrast, states that listeners only normalize to the extent that they know a language. Such an interpretation suggests that compensation is influenced by higher-level cognitive processes and relies on the listener's ability to

recognize words and phonemes in the precursor sentences. To be able to test this hypothesis more thoroughly, we compared perceptual adaptation patterns between native Spanish speakers with low proficiency in English (Spanish) and native Spanish speakers with high proficiency in English (Spanish-English bilinguals).

The third hypothesis states that the amount of compensation depends on the information value of F_1 in a listener's native language. Native speakers of Spanish, native speakers of English and native speakers of Dutch all identify vowels on the basis of F_1 and F_2. Listeners of English and Dutch, however, additionally rely on duration cues and diphthongization. A vowel's exact F_1- F_2 combination is thus more important for a native speaker of Spanish than a native speaker of English. If the amount of compensation depends on the information value of F_1, it is expected that native speakers of Spanish show the largest amount of compensation. Importantly, however, this hypothesis also predicts that when speakers of these three languages categorize sounds of a continuum solely determined by F_1, the native Spanish speakers will show the most reliable categorization (observed as the steepest categorization curve). This hypothesis would be further supported if, between participants, the size of a participant's compensation effect correlates with the steepness of a participant's categorization curve.

The final hypothesis states that perceptual compensation processes are shaped by the sound properties of the ambient language. This would suggest that native language learners who are exposed to a language with fewer vowel categories, that is, vowel categories that are positioned further apart in the vowel space, will have less of a need to perceptually compensate. Spanish has only five monophthong vowel categories whereas English and Dutch have 11 and 13 respectively. The exact number, though, varies across dialects (Gussenhoven, 1999; Ladefoged, 1999). Note that Spanish also has a number of diphthongs (Aguilar, 1999). However, the main cues for diphthongs are formant movement rather than steady F_1 and F_2 values. According to this hypothesis, native speakers of Spanish will normalize less than native speakers of Dutch and native speakers of English. This hypothesis, however, rests on the assumption that the vowel categories in Spanish not only lie far apart in vowel space, but also that they have similar within-category variance to those in English and Dutch. If within-category variance in Spanish is much higher than in English and Dutch, the overlap between phoneme categories could still be similar among these languages. In order to examine these properties of the vowel spaces

across languages, we collected vowel production data from Spanish, English and Dutch speakers in their native language. We measured the distance between categories and the variance within categories for the five vowels /i/, /e/, /a/, /o/ and /u/.

The goal of the current study was to investigate compensation for speaker vocal tract properties in listeners from different language backgrounds. Native speakers of Spanish, native speakers of English, native speakers of Dutch and Spanish-English bilinguals were tested with materials in all three languages. They categorized target sounds of a [sufu] to [sofo] continuum which were preceded by sentences that were manipulated to have either a high F_1 contour or a low F_1 contour. This experiment allowed us to test whether the strength of compensation processes is fully determined by signal properties of precursor-target combinations or whether language background has an additional influence.

Method

Participants.

Eighteen native listeners of American English (2 male, 16 female), Dutch (4 male, 14 female), Spanish (4 male, 14 female) and eighteen Spanish - English bilinguals (5 male, 13 female) participated in the study. The native English and Spanish-English bilingual participants were recruited and tested in the Phonetics Laboratory at the University of Texas - Austin. The native speakers of Spanish (low English proficiency) were recruited and tested on location at ESL schools in the Austin area and at the Phonetics Laboratory at UT. The native speakers of Dutch were recruited and tested at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands.

Each participant filled a detailed background language questionnaire that was adapted from the LEAP-Q questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007). Additional questions regarding the participants' dialectal background were included (data on these additional questions will not be addressed here). In order to encourage participants to engage in native-language speaking mode, each participant was interviewed prior to the experiment and instructed about the task in their native language by a native speaker.

Table 1 provides a summary of a number of measures regarding the language experience for all participants. The results indicate that the four groups of participants were qualitatively different from each other with respect to their dominant language, age of acquisition, the amount of exposure and their proficiency across the three

languages. This confirmed our initial assumptions about the differences in the language experience across our listener groups.

Table 1: Average age and language background information for all listeners. 'Dominant language' and 'Acquired first' reflect the percentage of participants that indicated that this language is the most dominant/first acquired language. 'Exposure' is calculated as the average amount of exposure to a language for the participants at this point in their lives. Because responses were unrestricted not all columns may add up to 100 (i.e., participants could indicate that they were also exposed to other languages). 'Proficiency' is calculated as the average over the self-rated proficiency in writing, speaking and understanding speech.

	Background			
Measure	Spanish	Bilingual	English	Dutch
Age (yrs)	34.6	21.9	20.8	21.6
Dominant language				
Spanish	100	39	0	0
English	0	61	100	0
Dutch	0	0	0	100
Acquired first				
Spanish	100	83	0	0
English	0	17	89	6
Dutch	0	0	0	94
Exposure				
Spanish	59	32	7	0
English	35	67	86	17
Dutch	0	0	0	77
Proficiency				
Spanish	100	86	15	0
English	54	96	100	76
Dutch	0	0	0	100

Production.

The participants produced the vowels /i/, /e/, /a/, /o/ and /u/ in /sVsV/ and /fVfV/ contexts (where V represents the target vowel). They were asked to produce them in their native language, and the Spanish-English bilinguals were asked to produce them in Spanish. The stimuli were presented to the speakers as a list on a computer screen. Participants read the list of the target non-words three times. Recordings were made using an Audio-technica AT2020 cardoid condenser microphone connected to a TASCAM US-144mkII usb audio interface. The signal was recorded onto a laptop using Adobe Audition software.

The recordings were inspected and a 40 ms portion from the most steady part (for both F_1 and F_2) close to the midpoint of the first vowel was selected. Note that this is not trivial, especially for diphthongized vowels in English. In case no "most steady part" could be found, the midpoint was taken. Five formants were estimated between 0 and 5300 Hz using Burg's formant estimation procedure in Praat (Boersma & Weenink, 2005). Measurements for the first two formants were used for further analyses.

Perception.

Recordings & Measurements.

For every language four carrier sentences were constructed that did not contain the critical sounds /u/ and /o/. These sentences also contained a limited number of sonorant phonemes because these give rise to more artifacts after the resynthesis method (described below). For every language one sentence was selected, based on whether it sounded natural after resynthesis. For Dutch this was an instance of the sentence "*Die kaas staat niet bij de*", 'That cheese does not stand with the', for Spanish this was "*a veces se halla*" 'at times she feels' and for English this was "*He loves eating fresh*". The target words were /sofo/ and /sufu/. These sentences were of similar length (i.e., 5 or 6 syllables).

One speaker of each of the three languages was recorded speaking the precursor sentences selected for their native language followed by the target non-words. The Spanish speaker was a speaker of Colombian Spanish, the Dutch speaker a speaker of standard Dutch, and the native English speaker spoke Midwestern Dialect of American English. We obtained a total of 24 instances of both /sufu/ and /sofo/ from each of the three speakers. Measurements of the duration of the target vowels and the trajectories of the first and second formant were taken.

Stimulus manipulation.

Targets.

The vowels were visually inspected and cut out of their fricative context at a zero-crossing. The position was selected as the transition point from the frication into the vowel or vice versa. The tokens were equalized in duration across the three speakers by cutting out individual periods (vowel 1: 147 ms; vowel 2: 165 ms). Using Burg's Formant method as implemented in Praat the filter characteristics of the targets were estimated. A source model was estimated with Burg's LPC method, using 80 predictors and thereby leaving little remnants of the formants. One source model of each of the two vowels in [sofo] was selected for each speaker. The onsets and offsets of the source models were ramped in amplitude to correct for differences in the location of zero crossings across tokens.

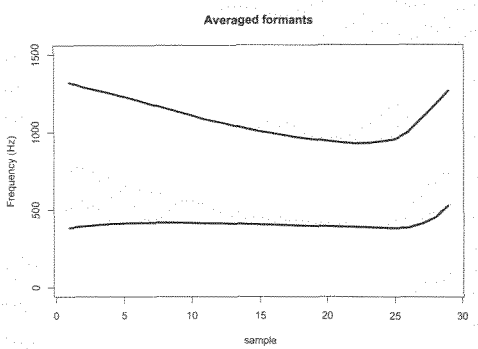

Figure 1: Averaged formant contours for the first and second formant of vowels that were recorded in a sVfV context (only the contours of the first vowel are displayed). The contours represent averages over a number of repetitions of the vowels [o] and [u], over the three different speakers that recorded the stimuli (who had either Spanish, English or Dutch as their native languages). These contours were used to resynthesize the target vowels for the three speakers, so that their formant tracks would be identical.

The filter model was estimated for all vowels by assuming two formants in the range between 0 and 2000 Hz. The trajectory of the F_1 and F_2 were estimated at around 30 points within each vowel (29 for the first vowel and 31 for the second). An average F_1 and F_2 track was calculated over productions of /o/ and /u/ in both the first and second position resulting in a single ambiguous filter model for the first vowel and one for the second vowel (i.e., with values halfway those in the average [o] and [u], for both formant height and formant bandwidth). This single average was based

on a number of repetitions of each of the three speakers. This dynamic filter therefore represented an average of the formant properties of over a number of instances of both [o] and [u], and averaged over the three speakers. Figure 1 displays the average F_1 and F_2 tracks for the initial target vowel. This procedure was followed so that the final manipulated target words had the same filter properties across the three different speakers. This assured that listeners could only rely on the same cues in the targets across the stimulus languages. In steps of 10 Hz, the height of only the first formant of the filter model was increased over a range of 100 Hz and decreased over a range of 100 Hz across the whole vowel to create the new formant models for the continuum from [u] to [o] (now only distinguished by F_1). The resulting formant models were then reapplied to the source models of a single instance of the first and second vowel of [sofo] for every speaker. So for each speaker the average filter properties were used but they were combined with a speaker-specific source model. The resulting signals were low-pass filtered between 0 and 1500 Hz (with the standard filter function in Praat that filters in the frequency domain with a smoothing of 100 Hz). These sounds were adjusted to have the same amplitude trajectory and overall amplitude as a low-pass filtered instance of the original vowel that was used for the source model. These sounds were then added to the high pass filtered parts (1500 - 6000 Hz) of the original vowels that had been used to create the source models (through summation of the signals across time). Vowels created in this way were spliced into the consonantal contexts of the particular speaker (the consonants were also equalized in duration across the speakers). All created items were equalized in amplitude. This yielded a total of 21 tokens of the [sofo] to [sufu] continuum per language/speaker.

We determined the 50% categorization-crossover points of the two vowels through pilot testing. Listeners were asked to categorize the targets of each of the three languages in a forced choice task from /sofo/, /sufu/, /sofu/ and /sufo/ response options. Based on this pretest, a 7-step range of vowels that covered an F_1 range of 120 Hz was selected. Each step was 20 Hz, ranging from, on average, 337 to 457 Hz for the first vowel and from 359 to 479 Hz for the second vowel. The top panels of Figure 2 display the LTAS of the endpoint vowels (those in initial position). The LTAS were calculated in Praat with a 10Hz bin-width. Note that the x-axis is logarithmic. The left-hand panel of Figure 3 displays the differences between the LTAS of the endpoint target vowels ([o] - [u]) It can be observed that the differences

are reasonably similar across the stimulus languages, which is expected since the same formant filter model was used for the resynthesis of the vowels.

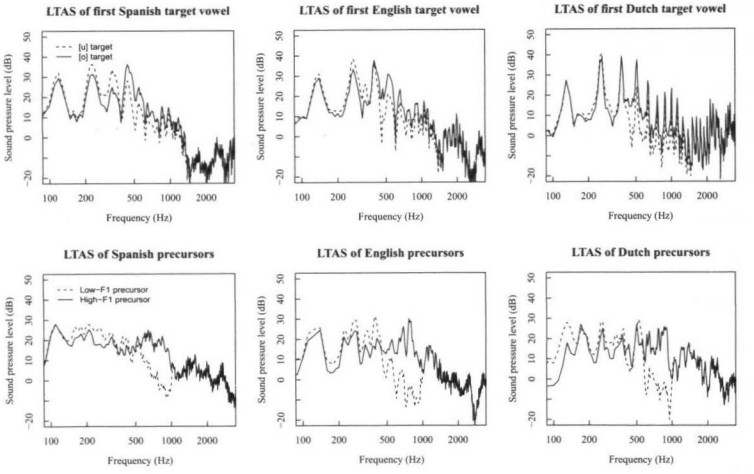

Figure 2: LTAS plots for the stimuli. Top panels: LTAS for the endpoint target vowels ([u] = dotted line, [o] = solid line) for, from left-to-right, the Spanish, English and the Dutch materials. Bottom panels: LTAS for the precursor sentences (Low F_1 = dotted line, High F_1 = solid line) for, from left-to-right, the Spanish, English and the Dutch materials. The x-axes are logarithmic.

Precursors.

The source and filter models of the precursors were again estimated with Burg's method in Praat, using the same parameters as for the targets. The F_1 track of the filter model was increased or decreased by 200 Hz (values are based on Watkins & Makin 1994) for each of the three speakers to create two new versions, one with a low and one with a high F_1 contour. One formant was estimated between 0 and 1000 Hz and the original signal was used for frequencies above 1000 Hz. This adaptation resulted in more natural sounding precursors. As with the targets, the precursor sound files were equalized for overall intensity and duration. Finally, the precursor sentences and targets were concatenated with a 500 ms silent interval between them. The relatively long precursor-target interval was used to prevent from peripheral auditory influences. The bottom panels of Figure 2 display the LTAS of the two contexts in each of the languages. The right-hand panel of Figure 3 displays the differences between the LTAS of the two context versions in each of the three languages.

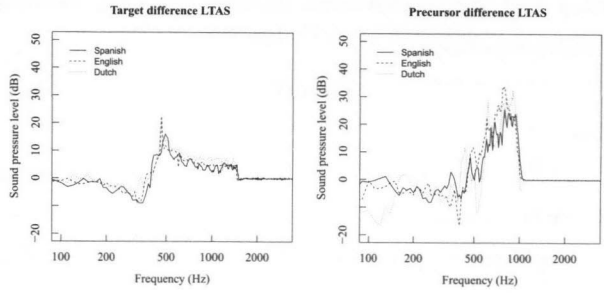

Figure 3: Difference LTAS lines. Left panel: difference LTAS for the endpoint targets for the three different language materials (Spanish = solid; English = dashed, Dutch = dotted). Right panel: difference LTAS for the precursors in the different stimulus languages (Spanish = solid; English = dashed, Dutch = dotted). The x-axes are logarithmic.

Although the differences between the different language materials are somewhat larger than for the targets, the differences are still reasonably similar. This indicates that, based purely on the LTAS, the effect sizes should be fairly similar across the stimulus sets.

Procedure.

After arriving to the testing site, the participants first filled out a language background questionnaire (LEAP-Q, Marian, Blumenfeld, & Kaushanskaya, 2007). Following that, they were recorded reading the list of non-words. Finally, they participated in the listening test. The different language-materials were presented in each of the six different possible orders to different participants following a Latin-square design. Each different ordering was presented to three participants of a particular language background. Listeners received written instructions about the task in their native language (a native speaker of their language was present to answer any additional questions). The participants were asked to categorize the last words of a stimulus as either /sufu/ or /sofo/. Listeners responded using the two buttons located below the touchpad on the laptop. The two options "sufu" and "sofo" were always displayed on the computer screen. A practice session preceded the test in order to familiarize the listeners with the task. Each of the 7 steps of the continuum was presented in both the low- and high- F_1 sentence conditions, randomly intermixed. Such a block was repeated 8 times per exposure language, resulting in 336 trials (2 precursor conditions * 7 steps * 8 repetitions* 3 languages). Stimuli were presented using Presentation

software (Version 11.3, Neurobehavioural Systems Inc.). The listening task took roughly 30 minutes per participant.

Results

Production data.

Figure 4 displays the average F_1 x F_2 values for each vowel for each group of speakers. Each symbol represents an average over approximately 108 vowel tokens (18 speakers * 2 non-words * ~3 repetitions; the number of repetitions is approximate as some instances, such as clear mispronunciations, were discarded). The two filled squares represent the two endpoint stimuli used in the perception experiment. Note that for the stimuli the formant values appear relatively low. This is due to the fact that the speakers for the stimuli were male, whereas most participants were female speakers, who tend to have higher formant values in general (Hillenbrand, Getty, Clark, & Wheeler, 1995). The figure shows that the five vowel categories occupy similar positions in the vowel space in all three languages. The exceptions are the English /o/ and /u/ which have a relatively high F_2 value, that is, they are fronted. This is a known phenomenon, especially in the Southern varieties of American English (AE) (Clopper, Pisoni, & de Jong, 2005).

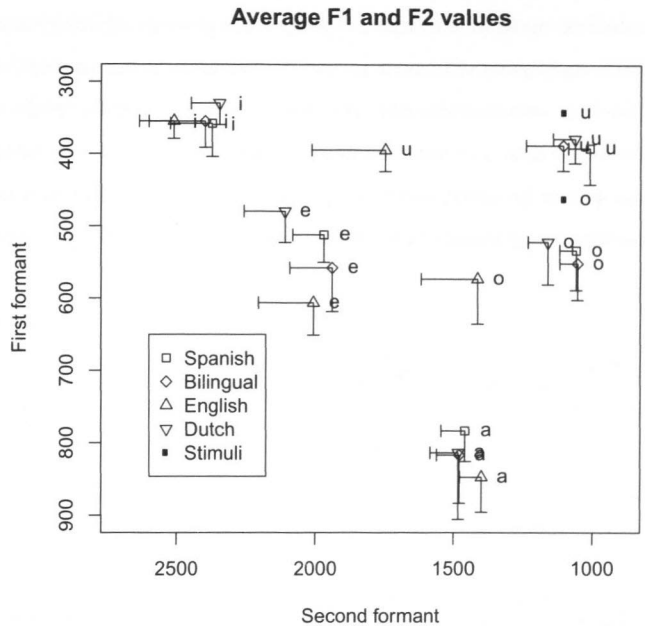

Figure 4: Average formant values plotted in F_1 - F_2 space, for four groups of speakers (Spanish, bilinguals, English and Dutch), each with18 speakers. The target endpoint values of the first vowel for the range of stimuli used in the listening experiment are also displayed. Participants were instructed to produce vowels in their native language (bilinguals were asked to say them in Spanish) in /sVsV/ and /fVfV/ context, where V represents the vowel. Measures are based on a 40 ms window in the steady part of the first vowel. Each participant repeated the items 3 times (in both /sVsV/ and /fVfV/ frames). Whiskers represent one standard deviation and are based on the same data as reported in Table 2.

Table 2: Between-speaker variation of the F_1 and F_2 in production of /i/, /e/, /a/, /o/ and /u/ for speakers of four different language backgrounds. Participants uttered /sVsV/ and /fVfV/ non-word sequences (where "V" represents the vowel). Measures are based on the first vowel. Measures were taken of the F_1 and F_2 values in those productions. Reported data represent the between-speaker variance, measured as the standard deviation over 18 participants per group. Values per participant were based on an average over ~3 repetitions of each non-word.

		Background			
Measure	Vowel	Spanish	Bilingual	English	Dutch
F_1					
	/i/	45.9	36.2	24.2	29.8
	/e/	38.1	60.5	44.7	43.2
	/a/	42.4	66.7	48.4	92.1
	/o/	54.9	50.5	62.2	58.8
	/u/	50.3	35.1	29.6	33.6
	Average	46.3	49.8	41.8	51.5
F_2					
	/i/	150.7	202.6	127.6	102.4
	/e/	113.5	154.1	198.9	145.7
	/a/	87.0	82.4	77.1	102.2
	/o/	60.0	64.9	203.4	72.5
	/u/	77.2	135.0	264.4	79.0
	Average	97.7	127.8	174.2	100.4

Note: for listeners of the same language background, standard deviations were calculated separately for speaker of different gender. These were then combined by calculating a weighted average.

Table 2 reports the between-participant variance of F_1 and F_2 in productions of /i/, /e/, /a/, /o/ and /u/ for speakers of the different language backgrounds. The results showed that the productions of the speakers from the different groups displayed similar amounts of variance. These results provide evidence that the overlap across various vowel categories in the F_1- F_2 space is a less severe problem in Spanish (because the five plotted vowels constitute the full Spanish monophthong inventory) compared to Dutch and English (which both have additional vowels lying between the plotted vowels).

Perception.

The data were analyzed using linear mixed-effects models in R (version 2.6.2, R development core team, 2008, with the lmer function from the lme4 package of Bates and Sarkar, 2007). Categorization responses were modeled using the logit-linking function (Dixon, 2008). Different models were tested in a deductive way, starting from a complete model including the factors Step (7 levels, ranging from -3 = [sufu] to 3 = [sofo], in steps of 1), Context (two levels, -1 = low- F_1 context and 1 = high- F_1 context) and Language (a categorical variable with the three levels: Spanish, English and Dutch). For the latter, Spanish served as the reference level. By-participant adjustments for the intercept, effects for Step and Context were included in the model (Baayen, Davidson, & Bates, 2008). This creates a better fit to the data. It was always tested whether the inclusion of these adjustments was justified by means of a likelihood ratio test, testing a model with these effects against a model without these by-participant adjustments with the anova() function in R (Baayen, et al., 2008). All numerical factors were centered around zero to make the interpretation of effects more straightforward (Barr, 2008). Non-significant predictors were taken out of the analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was also found. The best of these models was established by means of a likelihood ratio test. Additionally, the values for the by-participant adjustments for the slope of the categorization function and the size of the context effect were extracted from the models. These were used to measure correlations between the context effect and the slope and correlations between the context effect and the F_1-production measure from the production data (the difference in F_1 between instances of [o] and [u] for a particular participant).

Figure 5 displays the categorization results. Separate panels display the separate stimulus-language sets over the listener language-background groups. Each row shows the results from a different group of listeners (from top to bottom: Spanish, bilingual, English and Dutch listeners). The different columns display the stimulus language (from left to right: Spanish, English and Dutch materials). Each panel displays two categorization functions along the [sufu] to [sofo] continuum for stimuli that were presented in the context of a high- F_1 precursor sentence (dotted line) and a low- F_1 precursor sentence (solid line). A first set of analyses investigated the data for the different language background groups separately. A subsequent analysis investigated the data across the different language background groups to establish whether there were differences among those groups.

Spanish

The top three graphs in Figure 5 display the categorization data for the Spanish speakers in the three different language conditions. When listening to the Spanish materials (top row, first column) the Spanish participants gave overall more /o/ responses than /u/ responses ($b_{intercept} = 0.947$, $z = 5.524$, $p < 0.001$), they gave more /o/ responses towards the [o] end of the continuum ($b_{Step} = 0.907$, $z = 10.977$, $p < 0.001$) and they were significantly influenced by the F_1 range in the precursor sentence ($b_{Context} = -1.396$, $z = -6.531$, $p < 0.001$). The latter effect was in the expected, compensatory direction, that is, participants gave more /o/ responses (/o/ has a high F_1) in the context of a precursor with a low F_1 than in the context of a precursor with a high F_1. Furthermore, the effect of Context was smaller towards the [o] end of the continuum ($b_{Step*Context} = 0.120$, $z = 2.826$, $p = 0.005$), which can be seen by the lines for the two context conditions coming closer together at the [o] end of the continuum.

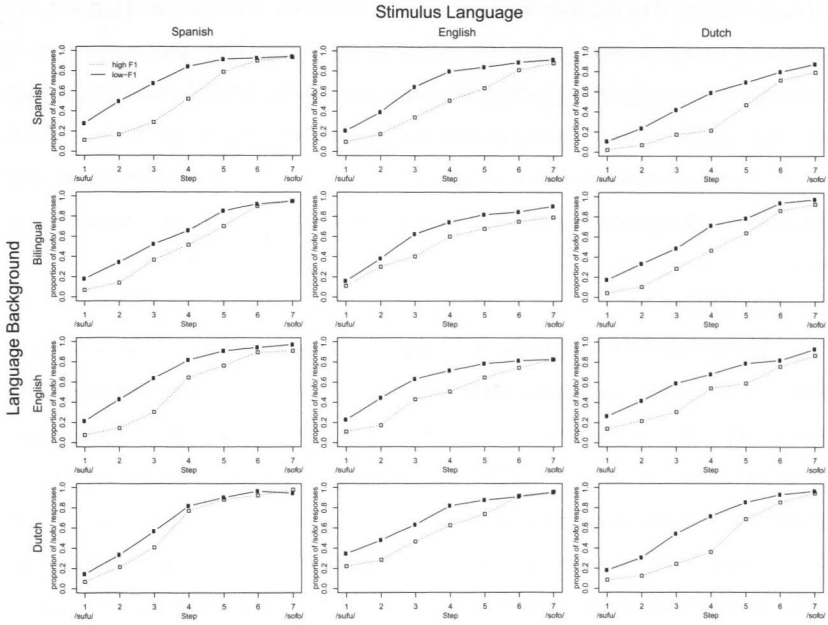

Figure 5: Categorization curves representing the probability of a /sofo/ response to targets of a 7 step [sufu] to [sofo] continuum. Targets were presented in the context of a precursor sentence that was manipulated to have either a generally low F_1 (solid line) or a generally high F_1 (dotted line). The stimuli were spoken in one of three possible languages, presented in separate blocks. The different stimulus languages are separately displayed on subsequent columns. From left-to-right the stimuli were spoken in: Spanish, English and Dutch. These stimuli were presented to four groups of 18 listeners with different language backgrounds. These are separately displayed on subsequent rows. From top-to-bottom the listeners had a Spanish, Spanish-English bilingual, English and Dutch language background.

When these data are compared to the condition where participants were listening to the English materials (top row, second column) it can be observed that with the English materials the /o/ bias was less pronounced, that is, more /u/ responses were given with the English materials than with the Spanish materials ($b_{Language(English)} = -0.343$, $z = -4.041$, $p < 0.001$). When listening to the Dutch materials (top row, third column) this bias was even more towards [u] ($b_{Language(Dutch)} = -1.320$, $z = -15.133$, $p < 0.001$). No other differences were observed. Compensation effects were found with all stimulus languages.

We also examined correlations between the size of the estimated effect of Context that a participant displayed with the F_1 distance between a participant's [o]

and [u] as measured in the non-word production task and the estimated steepness of a participant's categorization curve for Spanish listeners. Only one correlation is significant. For the Spanish participants listening to the English materials a just-significant correlation was found between Context and Slope (r = -0.469, t = -2.126, df = 16, p-value = 0.050), which suggests that in this case, those participants that showed steep categorization curves also showed a slightly stronger compensation effect.

Bilinguals

The second row of panels in Figure 5 displays the data for the bilingual listeners in the three language conditions. When these participants listened to the Spanish materials (second row, first column) they gave overall more /o/ responses than /u/ responses ($b_{intercept}$ = 0.594, z = 3.415, p < 0.001), they gave more /o/ responses towards the [o] end of the continuum (b_{Step} = 0.960, z = 14.135, p < 0.001) and their categorization responses were significantly influenced by the F_1 range of the speaker in the precursor sentence in the expected, contrastive, direction ($b_{Context}$ = -0.785, z = -3.588, p < 0.001). Just like the Spanish participants, the bilingual participants gave more /o/ responses in the context of a low- F_1 precursor than in the context of a high- F_1 precursor.

A comparison of the bilinguals' perception of the Spanish materials to their perception of the English materials (second row, second column) shows that with the English materials categorization responses were less categorical ($b_{Step*Language(English)}$ = -0.275, z = -5.631, p < 0.001), that is, the two lines are less steep. Unlike the bilinguals' data for the Spanish materials, their data on the English materials indicated that the effect of context was stronger towards the [o] end of the continuum ($b_{Step*Context*Language(English)}$ = -0.257, z = -2.627, p = 0.009). The lines are further apart towards the [o] end. When Spanish was compared with the Dutch materials, the /o/ bias was less strong for the Dutch materials ($b_{Language(Dutch)}$ = - 0.190, z = -2.139, p = 0.032). An increase in the effect of context fell just short of significance ($b_{Context*Language(Dutch)}$ = -0.336, z = -1.891, p = 0.060). Compensation effects were found with all stimulus languages. None of the correlations of the by participant adjustment of the effect of Context with the two relevant measures was significant. This means that there was no evidence that participants that had steep categorization curves also displayed stronger effects of context.

English

The third row of panels in Figure 5 displays the data for the English participants. When these participants listened to the Spanish materials (third row, first column) they displayed a strong /o/ bias ($b_{Intercept}$ = 0.956, z = 6.114, p < 0.001). Listeners gave more /o/ responses towards the [o] end of the continuum (b_{Step} = 1.026, z = 11.345, p < 0.001) and they were significantly influenced by the F_1 range of the speaker in the preceding sentence ($b_{Context}$ = -1.300, z = -5.320, p < 0.001), showing that the English participants also gave more /o/ responses in the context of a low- F_1 precursor than in the context of a high- F_1 precursor. The effect of Context was slightly stronger towards the [u] end of the continuum ($b_{Step*Context}$ = 0.081, z = 2.014, p = 0.044).

A comparison of the responses to the Spanish materials with the responses to the English materials shows that with the English materials the /o/ bias was less pronounced, that is, overall there were fewer /o/ responses ($b_{Language(English)}$ = -0.545, z = -6.257, p < 0.001). Responses were less categorical ($b_{Step*Language(English)}$ = -0.383, z = -7.513, p < 0.001) which can be observed by the fact that the curves are less steep. Further, the effect of the F_1 in the precursor sentence was somewhat smaller ($b_{Context*Language(English)}$ = 0.436, z = 2.555, p = 0.011).

A similar pattern was found when the responses to the Spanish and Dutch materials are compared. The /o/ bias was less pronounced ($b_{Language(Dutch)}$ = -0.518, z = -5.905, p < 0.001), listeners' responses were less categorical ($b_{Step*Language(Dutch)}$ = -0.327, z = -6.342, p < 0.001) and the effect of context was again somewhat smaller ($b_{Context*Language(Dutch)}$ = 0.436, z = 2.036, p = 0.042). Compensation effects were found with all stimulus languages. None of the correlations of Context with the two relevant by participant adjustments was significant.

Dutch

The bottom row of panels displays the results for the Dutch listeners. The leftmost panel displays the data for the Dutch listeners listening to the Spanish materials. The data show that listeners had a significant bias towards /o/ ($b_{Intercept}$ = 1.162, z = 6.386, p < 0.001). They gave more /o/ response towards the [o] end of the continuum (b_{Step} = 1.183, z = 12.354, p < 0.001) and listeners were again significantly influenced by the F_1 range of the speaker in the preceding context ($b_{Context}$ = -0.451, z = -2.465, p = 0.014). The effect of Context was smaller towards the [o] end of the continuum ($b_{Step*Context}$ = 0.120, z = 2.566, p = 0.010). The lines are further apart

towards the [u] end of the continuum.

Relative to the Spanish materials, the Dutch listeners' responses to the English materials were less categorical ($b_{Step*Language(English)} = -0.374$, $z = -6.533$, $p < 0.001$). The Dutch listeners' responses to the Dutch materials indicate that the number of /o/ responses was lower ($b_{Language(Dutch)} = -0.741$, $z = -7.613$, $p < 0.001$). Responses were less categorical ($b_{Step* Language(Dutch)} = -0.144$, $z = -2.423$, $p = 0.015$) and the effect of context was stronger with the Dutch materials than with the Spanish materials ($b_{Context* Language(Dutch)} = -0.816$, $z = -4.415$, $p < 0.001$). Compensation effects were found with all stimulus languages. None of the correlations of the by participant adjustment of the effect of Context with the two relevant measures was significant.

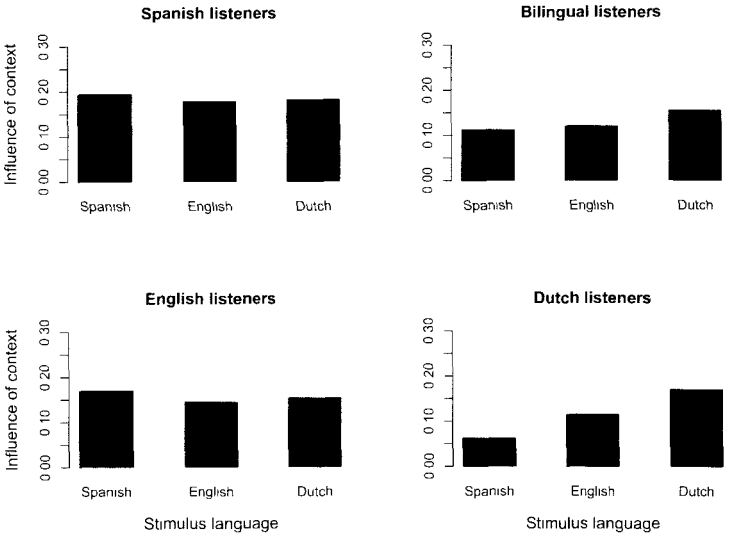

Figure 6: Bar-graphs representing the average difference between the high- F_1 and the low- F_1 categorization curves (that are themselves displayed in Figure 5). The bars represent the numerical size of the compensation effect. Each panel displays the effects for the three stimulus languages Spanish, English and Dutch. Separate panels display the effect for the groups of participants that had a different language background. From left-to-right and from top-to-bottom the listeners had a Spanish, bilingual, English or a Dutch language background.

Comparison between listener groups

Additionally, a comparison was made between the different groups of participants. Figure 6 displays the numerical effect size of the factor Context per condition. This analysis was based on the same factors as those for the analyses

described above with the addition of the factor Background (a categorical variable with the levels: Spanish, Bilinguals, English, Dutch). The data of the Spanish listeners listening to the Spanish materials serves as the reference level. Appendix A reports the values for the significant effects in the final model. Only the results that involve the interaction between the context effect with the factor Background will be discussed here. The optimal model showed that, compared to Spanish listeners listening to Spanish materials, both Dutch and bilingual listeners compensated less for the F_1 range in the preceding sentence when listening to Spanish materials ($b_{Context*Background(Bilingual)} = 0.736$, $z = 2.376$, $p = 0.018$; $b_{Context*Background(Dutch)} = 1.045$, $z = 3.332$, $p < 0.001$). Dutch listeners compensated more for the F_1 in a preceding sentence when listening to the Dutch or English materials than when listening to Spanish materials ($b_{Context*Background(Dutch)*Language(Dutch)} = -0.935$, $z = -3.659$, $p < 0.001$; $b_{Context*Background(Dutch)*Language(English)} = -0.532$, $z = -2.139$, $p = 0.033$).

A comparison between the optimal model and a model that included the same factors except for the interaction between Context by Background by Language showed that the critical interaction could not be dropped without significant loss of fit (anova() function in R: $\chi^2 = 24.909$, $df = 6$, $p < 0.001$). This shows that the effect of Context was significantly modulated by the interplay between the stimulus language and a participant's language background.

Summary

The focus of this study was to examine whether language experience modulates the influence of the height of F_1 in a precursor sentence on target vowel categorizations across different stimulus languages. Compensatory effects were observed in all listener groups and across all stimulus languages. Differences in the amount of compensation were observed between listener groups for some stimulus languages. Spanish listeners showed large compensation effects for all stimulus languages. The bilinguals, however, compensated less with the Spanish materials than the Spanish listeners did. Furthermore, bilingual listeners showed a tendency to compensate more for Dutch than other materials, but this effect fell just short of significance and was present only in the analysis of the bilingual participants. English listeners showed more compensation for the Spanish materials than for the English and Dutch materials (although, again, this effect was only significant in the separate analysis of the English listeners). The Dutch listeners compensated less for the Spanish materials than the Spanish listeners did. Moreover, the Dutch listeners

showed stronger compensation effects for the Dutch materials than for the Spanish materials. For the latter comparison the estimated b value for the context effect was roughly 2.5 times larger for the Dutch materials than for the Spanish materials. These differences show that the amount to which listeners compensate depends on their specific language background and the stimulus language they are listening to.

Some of the effects of the factor Context were significant in the analyses performed on the separate language background groups, but were not significant in the final overall analysis that included all listeners. The pattern of compensation across the different listener groups and sets of materials was not uniform. But the reliability of the differences in the size of the context effect across those sets could be called into question. The model comparison, however, showed that the inclusion of the critical Context by Background by Language interaction was justified.

Overall, the results also showed a bias towards /o/ responses. This effect was most pronounced for the Spanish materials and smaller for the English and Dutch materials. This suggests that, despite the pretest, the target stimuli were not completely perceptually balanced. Secondly, all listener groups were least categorical for the English materials. This effect could be due to the fact that the target test stimuli on the [o] - [u] continuum were least close to the standard AE pronunciation (recall that duration and formant tracks were equalized across the different language stimuli).

General Discussion

The current paper investigated the effect of language experience on perceptual accommodation to speaker vocal tract variation. Listeners categorized vowel targets on a [sufu] to [sofo] continuum. These targets were preceded by precursor sentences that had either a high or a low F_1 (following Ladefoged and Broadbent, 1957). Listeners from four language backgrounds participated in this study. These were native speakers of Spanish, Spanish-English bilinguals, native speakers of English and native speakers of Dutch. All participants listened to stimuli in Spanish, English and Dutch. This design allowed us to investigate language-dependent influences on perceptual compensation for speaker vocal tract characteristics.

The current study reveals a number of important findings. Related to our central question about the role of the language experience on perceptual adaptation, the results showed that listeners from all language backgrounds compensated for the vocal tract properties of a speaker in a precursor sentence. While previous reports have documented compensation effects in Dutch (Mitterer, 2006a; Sjerps, et al., 2011;

van Bergem, et al., 1988) and English (Ladefoged & Broadbent, 1957; Watkins, 1991; Watkins & Makin, 1996), the current study extends this effect to a Romance language with different vowel properties. Compensation for vocal tract characteristics thus seems to be a general property of listening to speech. Furthermore, listeners compensated for precursors when they listened to the materials in their own language as well as when they listened to the materials in their second language and in an unfamiliar language. This finding further emphasizes the generality of the mechanisms that cause compensation for vocal tract characteristics.

One of the most critical findings of this study, however, is that the perceptual impact of a precursor sentence on subsequent targets can vary. Specific combinations of language background and stimulus language led to stronger or weaker compensation effects. The fact that listeners from different language backgrounds are sensitive to different aspects of the precursor signals seems to have caused them to be differentially influenced when perceiving a subsequent target sound. Differences in the amount of compensation have been shown to occur across different types of manipulated speech sounds (Sjerps, et al., 2011). For instance, when a speech precursor was altered by manipulating a number of specific prosodic properties, the size of the compensation effect was reduced. This was despite the fact that the LTAS relations between precursor and targets remained similar. Those manipulations consisted of removal of pitch movement and silent intervals (due to stop closures) and the absence of natural attack and decay structures because syllables were temporally reversed (i.e., the first digital samples of the precursors traded places with the last samples, etcetera). The current study, however, shows that differences in the amount of compensation can also be found for the same speech stimuli, as a function of the language background of the listener. This latter result thus extends the finding of Sjerps et al. (2011) that compensation effects can vary in magnitude when LTAS relations between precursor and target are held constant. Identical precursor-target combinations can lead to different effects as a function of listeners acquired sensitivity.

At the outset of this research we had formulated four hypotheses. With regard to the hypothesis that Spanish listeners will show a smaller adaptation effect due to less overlap between the vowel categories, the production results revealed that between-speaker variance was similar across the four groups of participants. English and Dutch have a number of additional vowels in between the vowels that are shared

with Spanish. This suggested that category overlap among monophthong vowels is indeed a less severe problem in Spanish. If compensation processes were mostly driven by the properties of the ambient language, native speakers of Spanish would be expected to show the least compensation of the listener groups. In contrast to this, the perception experiment showed that Spanish listeners compensated for the F_1 in a preceding context to the largest extent of all listener groups. The current findings therefore do not provide support for vowel confusability in the ambient language as the explaining factor.

The next possible account of the observed perceptual adaptation patterns relates to language familiarity. The amount of compensation observed for the Dutch listeners seems to be related to their familiarity with the stimulus language. They compensated least when listening to the Spanish materials, somewhat more when they listened to the English materials (Dutch students all learn English in school) and most when listening to the materials of their native language. However, Spanish listeners, although they were low in proficiency in English and completely unfamiliar with Dutch, compensated to the same extent in all three languages. The bilingual listeners showed a trend towards stronger compensation effects when listening to Dutch than when listening to Spanish. Furthermore, the analysis on only the English listeners indicated that the English listeners compensated more strongly with the Spanish than with the English materials (although this effect was not reflected in the overall analysis, indicating that the effect is not very strong). Combined, these results suggest that compensation effects may in part be shaped by the listener's experience with language, but this effect does not directly reflect familiarity to a particular language per se.

A more basic assumption was that compensation is completely independent of specific influences of language background and stimulus languages. This assumption is in accordance with the fact that compensation effects were found for all listener groups across all stimulus sets. There are a number of findings in this study, however, that do not fully support this most basic hypothesis. When comparing the responses to the Spanish materials across the listener groups, both the bilingual and the Dutch listeners show significantly less compensation than the native speakers of Spanish. The results in the current study therefore do not support the assumption that the relations between signal properties of the precursors and targets can fully explain the effect that a precursor has on a target. An additional test confirmed that the inclusion

of the Background by Language by Context interaction was indeed justified. It is unlikely that rejecting this assumption is based on a spurious effect.

Finally, unlike listeners of English and Dutch, listeners of Spanish may rely more heavily on formant values than on duration when recognizing vowels. This may account for the Spanish listeners' strongest compensation effects. The impact of the F_1 contour is more important in target vowel judgments for native speakers of Spanish than for native speakers of English and native speakers of Dutch. It is thus possible that the F_1 contour in a preceding sentence will also have a stronger perceptual impact on such target judgments for native speakers of Spanish. This prediction is supported by the Spanish data. Spanish listeners appeared to apply their native perceptual strategies to an unfamiliar language, that is, they compensated strongly and to a similar extent across the three sets of materials. The bilingual listeners also compensated to a similar extent across the different sets of materials. They compensated less than the native speakers of Spanish in general though, but this could be due to their experience with English. The patterns for the native speaker of English and Dutch, however, are more difficult to reconcile with this approach. The English listeners did not compensate less with the Spanish materials than the Spanish did. The Dutch listeners compensated to a different extent across the materials. Furthermore, as indicated in the introduction, if the importance of F_1 were a driving factor behind the amount of compensation, then compensation should be stronger for individuals that have a steep categorization curve. Except for one analysis (the Spanish participants listening to the English materials), none of the tests revealed a correlation between the steepness of a participant's categorization curve and the size of the compensation effect. These combined findings do not support the interpretation that the importance of F_1 as a cue to vowel categorization fully determines the amount of compensation that is observed.

None of the outlined hypotheses were fully confirmed. However, a number of important conclusions can be drawn from this study. We observed an influence of precursors on subsequent targets for all listener groups and with all stimuli. However, we also found that signal properties of the precursor and targets do not fully predict the observed effect sizes. Listeners from different language backgrounds naturally pick up on different perceptual cues in speech signals. We show here that such different sensitivities also influence the strength of the impact that the LTAS of a precursor signal has on a subsequent target. It will be a future challenge to find out

what specific type of cues can lead to such an increase or decrease of precursor-LTAS impact.

Conclusion

Perceptual compensation processes aid listeners in dealing with across-talker variation in speech. Evidence is accumulating that these processes are based to a large degree on relatively general auditory processes (Sjerps, et al., 2011; Stilp, et al., 2010; Watkins, 1991; Watkins & Makin, 1994, 1996). In line with these findings the current study shows that compensation effects can be found with the same materials across three different languages and for listeners who are listening to a second language or an unfamiliar language. Such compensation mechanisms may increase the effective transfer of information in perception (Kiefte & Kluender, 2008; Kluender, et al., 2003; Kluender & Kiefte, 2006). The present study, however, shows that compensation for speaker vocal tract properties is a gradual process. This finding indicates that the impact of compensatory processes can be modified by factors such as a listener's language background and the language of the stimuli. The size of the compensation effect not only depends on the precise acoustic properties of the stimuli but also on a listener's previous exposure to language.

Appendix A

Variable	b	z	p
(Intercept)	0.992	5.854	< 0.001
Step	0.969	11.346	< 0.001
Context	-1.536	-6.916	< 0.001
$Language_{(Dutch)}$	-1.356	-15.084	< 0.001
$Language_{(English)}$	-0.400	-4.515	< 0.001
Step by Context	0.076	3.533	< 0.001
Context by $Background_{(Bilingual)}$	0.736	2.376	0.018
Context by $Background_{(Dutch)}$	1.045	3.332	< 0.001
Step by $Language_{(English)}$	-0.125	-2.406	0.016
$Background_{(Bilingual)}$ by $Language_{(Dutch)}$	1.165	9.305	< 0.001
$Background_{(Dutch)}$ by $Language_{(Dutch)}$	0.623	4.707	< 0.001
$Background_{(English)}$ by $Language_{(Dutch)}$	0.838	6.674	< 0.001
$Background_{(Bilingual)}$ by $Language_{(English)}$	0.314	2.586	0.010
$Background_{(Dutch)}$ by $Language_{(English)}$	0.295	2.241	0.025
Step by $Background_{(English)}$ by $Language_{(Dutch)}$	-0.262	-3.550	< 0.001
Step by $Background_{(Bilingual)}$ by $Language_{(English)}$	-0.157	-2.212	0.027
Step by $Background_{(Dutch)}$ by $Language_{(English)}$	-0.247	-3.208	0.001
Step by $Background_{(English)}$ by $Language_{(English)}$	-0.259	-3.563	< 0.001
Context by $Background_{(Dutch)}$ by $Language_{(Dutch)}$	-0.935	-3.659	< 0.001
Context by $Background_{(Dutch)}$ by $Language_{(English)}$	-0.532	-2.139	0.033

Evidence for pre-categorical extrinsic vowel normalization

Chapter 5

Sjerps, M. J., McQueen, J. M., and Mitterer, H. (under revision). Evidence for pre-categorical extrinsic vowel normalization. *Attention, Perception & Psychophysics*.

Abstract

Three experiments investigated the cognitive locus of extrinsic vowel normalization. Does normalization take place at a categorical or a pre-categorical level of processing? Traditional vowel normalization effects in categorization were replicated in Experiment 1. Vowels taken from an [ɪ]-[ɛ] continuum were more often interpreted as /ɪ/ (that has a low first formant) when preceded by a context that had a raised first formant (F_1), than if the context had a lowered F_1. This was established with a context that consisted of only two syllables. This short context was necessary for Experiment 2, a discrimination task that encouraged listeners to focus on perceptual properties of vowels at a pre-categorical level. Vowel normalization was again found: Ambiguous vowels were more easily discriminated from an endpoint [ɛ] than from an endpoint [ɪ] in a high F_1 context whereas the opposite was true in a low F_1 context. Experiment 3 measured discriminability in steps along the [ɪ]-[ɛ] continuum. Contextual influences were again found, but without discrimination peaks. The latter indicated that listeners had focused on pre-categorical properties of the stimuli. Extrinsic vowel normalization therefore appears to be a process that takes place at least in part at a pre-categorical processing level.

CHAPTER 5: PRE-CATEGORICAL EXTRINSIC NORMALIZATION

Introduction

When listening to speech, listeners are faced with the problem that any particular phoneme is never realized twice in exactly the same way. The production of a speech sound can vary due to factors such as a speaker's gender or accent, or due to the phonetic context in which a phoneme is uttered (Heinz & Stevens, 1961; Hillenbrand, Clark, & Nearey, 2001; Hillenbrand, et al., 1995; Purnell, Idsardi, & Baugh, 1999). The variation that arises as a result of these factors can be so severe that phonemes overlap with respect to their most important auditory cues (such as the first two formants, F_1 and F_2, in the case of vowels). Under normal listening conditions, however, listeners are hardly bothered by this variation. Part of the solution to this apparent contradiction lies in the fact that speech is perceived relative to general voice characteristics such as a speaker's pitch (Miller, 1953; Nearey, 1989) and higher formants (Nearey, 1989). The perception of vowels is influenced not only by vowel-intrinsic aspects (like pitch and higher formants) but also by vowel-extrinsic context (Ladefoged & Broadbent, 1957). Listeners interpret vowels relative to vocal tract characteristics that are revealed in a preceding sentence. If the speaker has a relatively high F_1, listeners interpret more ambiguous [ɪ]-[ɛ] sounds as /ɪ/ (the vowel with the relatively low F_1), whereas more vowels are interpreted as /ɛ/ when the speaker has a generally low F_1. This study investigates the cognitive locus of listeners' ability to use extrinsic information to compensate for a speaker's vocal tract characteristics.

Johnson et al. (1999) have demonstrated that listeners' categorizations of vowels can also be changed through more abstract knowledge such as perceived gender. They showed that categorization behavior of an F_1 [ʊ] - [ʌ] continuum differs depending on whether listeners saw a movie of a male or of a female speaker (females generally have higher F_1 values than males). Moreover, they also showed that a similar influence can be found when listeners are made to believe that they are listening to a female or a male speaker (through instructions). Such results suggest that listeners perceive vowels relative to a cognitive model of the expected vowel space of a particular speaker. Normalization effects would then be the result of a speaker-dependent restructuring of the cognitive vowel space. The idea of such a restructuring of category boundaries conflicts with another proposal about vowel normalization. This proposal suggests that extrinsic vowel normalization has a mainly auditory basis (Watkins, 1991) and takes place at an auditory level of processing. In

this approach, the normalization process has been formulated as a mechanism that changes the perception of vowels by taking the overall average spectral shape of a speaker's context sentence and inversely filtering new auditory input for that average (Watkins, 1991; Watkins & Makin, 1994). As a result the context-dependent adjustment of the auditory representation leads to context-dependent perception of phoneme identity.

This study investigated whether vowel-normalization effects like those reported by Ladefoged and Broadbent (1957) are best explained by a speaker-dependent restructuring of vowel space, or a change that takes place earlier in the processing hierarchy, at a level of more basic pre-categorical representations. A speaker dependent restructuring of vowel space would imply a context-dependent change in the location of phoneme category boundaries. A change in pre-categorical representations would imply a context-dependent change in the auditory coding preceding phoneme categories. If normalization is the result of a shift in category boundaries, this should not influence the auditory coding that precedes the activation of phonemic categories. The two hypotheses can therefore be compared by testing extrinsic normalization in a task that does not rely on the use of phoneme categories. This can be investigated by comparing discrimination behavior with categorization behavior (Clarke-Davidson, Luce, & Sawusch, 2008; Kingston & Macmillan, 1995; Mitterer, Csepe, & Blomert, 2006). There is an important prerequisite for being able to test this prediction however: Listeners in the discrimination task need to be focusing on the pre-categorical rather than the categorical aspects of the speech stimuli. If listeners' discrimination responses indeed reflect pre-categorical processing, then the category-shift account of normalization predicts that there will be no normalization effect because pre-categorical representations are immune to the influence of the categorical normalization mechanism. In contrast, the auditory account predicts that normalization effects will be detected in responses that reflect a pre-categorical level of processing.

We examined the locus of extrinsic vowel normalization by comparing listeners' performance on categorization and discrimination tasks using the same stimuli across tasks. Comparing categorization with discrimination behavior has a long history in research on the perception of speech sounds. Recently, Clarke-Davidson et al. (2008) relied on this method of comparison to argue against a bias interpretation of perceptual learning in speech perception. They showed that the shift

in categorization behavior that is traditionally found in perceptual learning research is accompanied by a shift in discrimination peaks (using a range of sounds from [s] to [ʃ]). They thus elaborated on the classic findings (Liberman, et al., 1957) based on the comparison of categorization and discrimination behavior. Liberman et al. (1957) showed that boundaries in categorization were accompanied by a peak in discrimination behavior (better discrimination at category boundaries) and argued that the discrimination peaks that appeared at phoneme category boundaries reflected the specialized processing of a speech perception module.

Gerrits and Schouten (2004) have shown, however, that the categorical nature of speech perception is related to the task that is used. They observed that discrimination, without discrimination peaks at category boundaries, can be found in some but not all discrimination tasks. Gerrits and Schouten investigated two different discrimination tasks: 2I2AFC (Two-Interval Two-Alternative Forced-Choice), in which on every trial a participant listens to two sounds and has to decide whether the order was AB or BA; and 4I-oddity (or 4I2AFC: Four-Interval Two-Alternative Forced-Choice), in which on every trial a participant listens to four sounds, containing three standards (S) and one deviant (D). The listener has to decide whether the order was SDSS or SSDS by indicating whether the deviant stimulus was in the second or third position. Gerrits and Schouten (2004) used the same stimuli in both tasks and found that 2I2AFC gives rise to a discrimination peak at category boundaries while 4I-oddity does not. They argued that 2I2AFC still induced categorical processing because listeners partly relied on phonetic labels. With 4I-oddity, however, listeners were encouraged to rely on an auditory level of representation.

It has also been found that consonantal stimuli more often give rise to discrimination peaks than vowels (Fry, et al., 1962). This is possibly a result of the fact that consonant information is transient and therefore listeners often only have their category labels left to rely on. For vowels, which are often longer and more stationary, it is easier to focus on auditory information because vowel cues are stretched over longer time spans (Pisoni, 1973, 1975). Discrimination peaks can nevertheless sometimes be found with vowels, however (Repp, Healy, & Crowder, 1979). Using a same-different (AX) task, Repp et al. (1979) showed that vowel discrimination performance was predicted quite well by vowel categorization behavior, but only if the category labels were obtained in response to the stimulus pairs used in the discrimination task. This suggests that performance on the AX

discrimination task, like the 2I2AFC task, may be mediated at least in part by phonetic category labels.

These prior findings suggest that using a 4I-oddity task with vowel stimuli, as opposed to AX or 2I2AFC tasks, will encourage participants to focus on auditory properties of the sounds. That was therefore the approach that was taken here. We tested for extrinsic vowel normalization on the same stimuli, using categorization and 4I-oddity discrimination tasks. It was predicted that, if extrinsic vowel normalization does not occur at a pre-categorical level of processing, and instead reflects only categorical processing, it would be reflected in categorization behavior only, and thus not in 4I-oddity discrimination behavior. If, in contrast, extrinsic vowel normalization does take place at a pre-categorical level, then vowel normalization should be found in both tasks.

The first experiment used a categorization task, the second experiment used a discrimination task. The same stimuli were used in both experiments. The 4I-oddity task that was used in Experiment 2 required relatively short stimuli because listeners had to make decisions about sets of four stimuli. To achieve this, listeners were tested on stimuli on the continuum from [ɪpapu] to [ɛpapu]. The [papu] part was manipulated to have either a generally high F_1 or a generally low F_1. Unlike most previous experiments on extrinsic vowel normalization, which use sentence-length contexts (Broadbent & Ladefoged, 1960; Darwin, McKeown, & Kirby, 1989; Ladefoged & Broadbent, 1957; Watkins, 1991; Watkins & Makin, 1994, 1996), relatively short stimuli have rarely been used. Verbrugge, Strange, Shankweiler and Edman (1976) did use short contexts but they found no significant changes in vowel identification. Experiment 1 thus sought to establish if the stimuli required for Experiment 2 could induce vowel normalization in a categorization task.

Within the 4I-oddity trials that were planned for Experiment 2, there would be a sequence of four trisyllables from the [ɪpapu] to [ɛpapu] continuum. The [papu] part following vowel x was intended to provide the preceding context for vowel x+1. This method was important because it made it possible to use a silent interval (500 ms) between the contexts and the subsequent target vowel while also providing every vowel with a similar amount of contextual influence. The silent interval is necessary because shorter intervals lead to increasing peripheral auditory influences (Summerfield, et al., 1984; Wilson, 1970) that are qualitatively indistinguishable from vowel normalization effects (see also Watkins, 1991). Additionally, this approach

made it necessary to present the different speaker conditions in separate blocks, so that the preceding context of the first trisyllable in a quadruplet would have the same preceding context as the second trisyllable. To assure that contextual influences of the [papu] part were similar across experiments the blocked design was also adopted in Experiment 1. The [papu] part in a categorization trial thus provided the context for the first vowel in the subsequent trial and was consistent over the course of a block.

Experiment 1 tested whether these materials and this design could elicit the traditional normalization finding. If these methods indeed resulted in categorization shifts, the same materials could be used in the 4I-oddity discrimination task in Experiment 2.

Experiment 1: Categorization

Method

Materials.

The stimuli consisted of the sequences [ɪpapu] ("ipapu") and [ɛpapu] ("epapu"), that are meaningless in Dutch. These nonsense words were spoken by a female native speaker of Dutch. The materials were further processed with Praat software (Boersma & Weenink, 2005). The first vowel of a recording of [ɛpapu] was excised. Based on Burg's Linear Predictive Coding (LPC) procedure in Praat a filter model was obtained for the vowel by estimating 4 formants between 0 and 5500 Hz. A source model was estimated with 8 prediction coefficients. A range of filter models spanning 200 Hz was created by a linear decrease of F_1 in steps of 1 Hz. These filter models were combined with the source model. The vowels were low-pass filtered between 0 and 1000 Hz and then combined with the higher frequencies of the original vowel (1000-5000 Hz) to make the manipulated stimuli sound natural. Filtering was conducted with the standard filter function in Praat that filters in the frequency domain (with a smoothing of 100 Hz). All manipulated vowels were adjusted so that their overall amplitude and their amplitude envelope matched those of the original vowel. From the target range ten target steps were selected ranging from [ɪ] to [ɛ]. These steps spanned an F_1 range of 180 Hz in steps of 20 Hz (with F_1 values ranging from 190 to 10 Hz lower than the recorded [ɛ]). The original [ɛ] had an F_1 value of 734 Hz (measured in the middle of the vowel). The F_1 of the created vowel continuum thus ranged from 544 Hz (step 1 represents [ɪ] which had an F_1 decrease of 190 Hz) to 724 Hz (step 10 represents [ɛ] which had an F_1 decrease of 10 Hz). The range of 180 Hz for F_1 that was used here is somewhat larger than the difference between /ɛ/ and /ɪ/

reported for females Northern Standard Dutch, which is estimated at 136 Hz (Adank, van Hout, et al., 2004).

The context part of the stimuli ([papu]) was manipulated by the same procedure but then with either an increase of F_1 or a decrease of F_1 by 200 Hz for both vowels. The original [papu] context vowels had F_1 values of approximately 730 Hz ([a]) and 410 Hz ([u]). The manipulated sounds from the [ɪ] to [ɛ] continuum were spliced onto the context [papu] parts to create the different steps on the [ɪpapu] to [ɛpapu] continuum. This resulted in 10 (steps) * 2 (contexts) = 20 different stimuli.

Design and Procedure.

Participants categorized all steps from the [ɪpapu] to [ɛpapu] continuum in both F_1-speaker conditions. The F_1-speaker conditions were presented in two separate parts with the order of presentation of those parts balanced over participants. The 10 steps from the continuum were randomly presented in each of 11 blocks within each F_1-speaker part. This resulted in a total of 220 trials per participant (10 targets * 11 repetition blocks * 2 F_1-speaker parts), interrupted once by a self-paced pause that separated the two F_1-speaker parts. During the experiment participants saw the labels "Ipapu" and "Epapu" on a computer screen in front of them. They were asked to identify each stimulus by clicking on the left or right mouse button. Participants were tested in a sound-proof booth, wearing Sennheiser HD 280-13 headphones. Stimuli were presented at a comfortable listening level. The experiment was run using Presentation software (Version 11.3, Neurobehavioural Systems Inc.).

Participants.

Ten participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. They received a monetary reward for their participation. None of the participants reported a hearing disorder, language impairment or uncorrected visual impairment.

Results

Responses faster than 100 ms after target onset were excluded (99.3% of the data were kept). The data were analyzed using linear mixed-effects models in R (version 2.6.2 R Development Core Team, 2008, with the lmer function from the lme4 package of Bates and Sarkar, 2007) Categorization responses were modeled using the logit-linking function (Dixon, 2008). Different models were tested in a deductive way, starting from a complete model including the factors Step (a 10-step continuum in steps of 1 ([ɪpapu] = -4.5; [ɛpapu] = 4.5)), Context (Low F_1 = -1 vs. High F_1 = 1),

Block (1 = -5; 11 = 5, in steps of 1), Order (first part = -1; second part = 1) and all their possible interactions. All factors were centered around zero to make the interpretation of effects more straightforward. Non-significant predictors were taken out of the analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was also found. The best of these models was established by means of a likelihood ratio test (with the *anova()* function in R).

Figure 1 shows the average categorization results. The optimal model for the data had significant main effects for the factor Step ($b = -0.782$, $p < 0.001$), that indicates the decrease of /ɪpapu/ responses towards the [ɛpapu] end of the continuum; Context ($b = 0.604$, $p < 0.001$), that reflects more /ɪpapu/ responses after a High- F_1 speaker context compared to the Low- F_1 speaker context; and Order ($b = 0.160$, $p = 0.017$), that reflects more /ɪpapu/ responses in the second half.

The model revealed two-way interactions between the factors Step and Order ($b = -0.126$, $p < 0.001$), that reflects the fact that participants were in general more categorical during the second half; Step and Context ($b = 0.104$, $p = 0.002$), reflecting that the effect of context is stronger towards the [ɛpapu] end of the continuum; and Context and Block ($b = -0.094$, $p < 0.001$) because the effect of Context became smaller as the experiment progressed.

Three-way interactions were found between the factors Step, Order and Context ($b = 0.116$, $p < 0.001$), if the Low F_1 condition was presented in the first half then the effect of context was relatively stronger on the [ɪ] side of the continuum; Step, Order and Block ($b = 0.046$, $p < 0.001$), because during the first half categorization behavior became increasingly more categorical whereas this increase had leveled off during the second half; Step, Context and Block ($b = 0.021$, $p = 0.035$), because for the Low-1 condition categorization became steeper during the experiment whereas this was not the case for the High-1 condition; and Order, Context and Block ($b = -0.061$, $p = 0.004$) because the Context * Block interaction was strongest in the second half). The overall pattern was that responses became more categorical over the experiment. The effect of context decreased over the experiment, but the effect did not reverse.

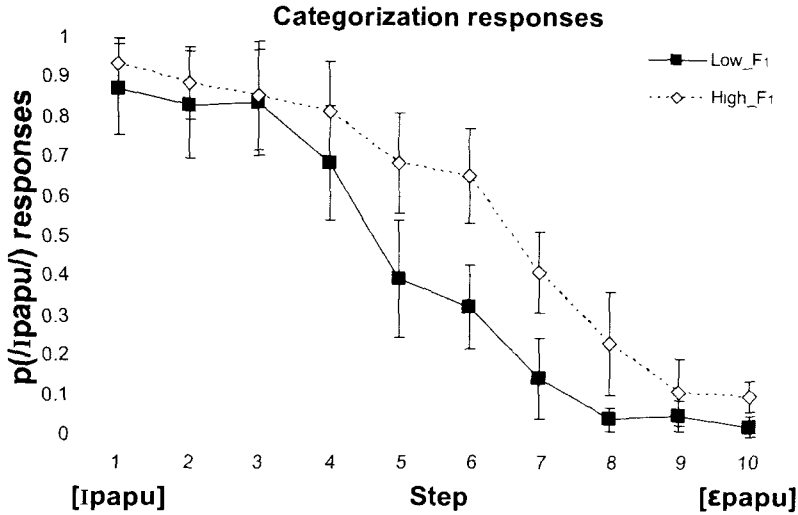

Figure 1. Experiment 1. Mean probability of an /ɪpapu/ response to a range of target sounds from [ɪpapu] (step 1) to [ɛpapu] (step 10). The [papu] part was manipulated to have an increased F_1 (High-F_1) or decreased F_1 (Low-F_1) level. Error bars reflect confidence intervals per step.

Discussion

In line with previous findings (Ladefoged & Broadbent, 1957; Watkins, 1991) compensation for speaker- F_1 characteristics was found with vowel targets on an [ɪ] to [ɛ] continuum. The vowels were presented in the context of a speaker with a high F_1 or a low F_1. Listeners categorized more sounds on the [ɪ] to [ɛ] continuum as /ɪ/ in the high- F_1 speaker condition than in the low- F_1 speaker condition. This effect was in the expected, compensatory, direction. This shows that a relatively short amount of speaker context (i.e., two syllables) can induce shifts in the perception of vowel identity in a compensatory manner. It also shows that the construction that was used here, where the [papu] part of item x influenced the interpretation of the initial vowel of x +1, was successful. It is important to note, however, that with this design it is uncertain whether the following context also influenced perception of the target vowels. Moreover, the fact that speaker conditions were presented in a blocked fashion might also have had an influence on the amount of normalization. These aspects, however, are not important for the investigation in this paper. The blocked presentation was necessary for the comparison of these results with the discrimination experiment.

Experiment 2 was set up to investigate whether these compensation effects would also be observed in a task that encouraged listeners to focus on the auditory aspects of the stimuli. The discrimination task that was used was the 4I-oddity task, in which listeners heard sets of three ambiguous standards (S) and one unambiguous deviant (D, either [ɪ] or [ɛ]), in one of two possible orders: SDSS or SSDS. In the High- F_1 speaker condition, an influence of the speaker context should be reflected in lower discriminability for [ɪ] from the ambiguous standards than for [ɛ] from the ambiguous stimuli. This prediction follows from the results from Experiment 1 (see Figure 1), that show that in the High- F_1 condition an ambiguous sound is perceived as more similar to [ɪ] than to [ɛ]. In the Low- F_1 [papu] context, however, it should be more difficult to discriminate [ɛ] from the ambiguous standards and easier to discriminate [ɪ] from the standards. Because there was a risk that listeners would reach ceiling performance or stay at floor performance (which could hide any possible context effects) the discrimination experiment was based on both larger (5 steps =100 Hz) and smaller (4 steps = 80 Hz) step sizes. A category shift-account of normalization predicts that the normalization effect arises at a categorical level of processing and not at an pre-categorical level of processing. This account predicts that no effect of normalization should be observed in Experiment 2. The auditory proposal assumes that effects of normalization take place at an auditory level and should be observed in the 4I-oddity procedure.

Experiment 2: Discrimination

Method

Materials and procedure.

Stimuli consisted of standard-deviant quadruplets in which the deviant was placed in second or third position, with 500 ms ISI between the individual trisyllabic sequences. The deviant consisted of one of the endpoints (step 1: [ɪ] or step 10: [ɛ]). The standard was a step from the middle of the continuum, either step 5 or step 6 (the frequency intervals between two members of a pair was thus either 80 or 100 Hz). These stimuli were created in both the High- F_1 and the Low- F_1 speaker conditions resulting in 2 (deviant is step 1 or 10) * 2 (standard is step 5 or 6) * 2 (F_1-speaker condition) = 8 types of stimuli.

Stimuli were presented in blocks of eight stimuli with the number of trials where the deviant would be in second or third position balanced (and the order was

randomized). Such a block was presented four times, resulting in 32 stimuli of one standard-deviant pair in a row. Stimuli from the eight different types of stimuli were presented as separate blocks that were presented in a random order. Participants responded by clicking on the left button of a button box (labeled "2": if they thought that the deviant was in second position) or the right button of a button box (labeled "3": if they thought that the deviant was in third position). Participants received visual feedback (printed "Correct" (correct) or "Fout"(incorrect) on the computer screen immediately after each response. This should provide participants with reinforcement on whether they were focusing on the right cues, and as such improve their overall discrimination performance. Note that, in Experiment 1 (which consisted of categorization) participants had not received feedback. We decided not to use feedback in categorization because feedback could introduce a bias on subsequent trials, and as such move the measured category boundary away from it's natural position.

Participants.

Eight participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. These were selected according to the same criteria as those for Experiment 1. They received a monetary reward for their participation.

Results

No participant responded earlier than 100ms after the onset of the second item. Figure 2 shows the average discrimination results. Analyses included the numeric factors Stepsize (Levels: -1 = 4 steps and 1 = 5 steps), Deviant (Levels: -1 = step 1, or [ɪ] and 1 = step 10, or [ɛ]) , Block (1 = -3.5, 8 = 3.5, in steps of 1) and Context (Levels -1 = Low F_1 and 1 = High F_1). The optimal model fitted to the data had significant main effects for the Intercept ($b = 2.110$, $p < 0.001$) reflecting overall performance above chance level; the factor Stepsize ($b = 0.453$, $p < 0.001$) that reflects an improvement in performance for larger step sizes; the factor Context ($b = -0.183$, $p = 0.024$) because performance was in general higher for the Low F_1 condition; and Block ($b = 0.112$, $p = 0.002$) that indicates that overall performance improved as the experiment progressed.

There were interaction effects between the factors Stepsize and Context ($b = 0.181$, $p = 0.024$) that indicate that the difference in discriminability between the two step sizes is larger for the High F_1 context stimuli than for the Low F_1 context stimuli;

Stepsize and Block (b = -0.171, p< 0.001) that indicates that the difference in discriminability between small and large steps became smaller as the experiment progressed; Deviant and Block (b = 0.121, p = 0.001) that reflects that the discrimination of [ɪpapu] became easier as the experiment progressed, whereas the discrimination of [ɛpapu] showed the opposite pattern; and Deviant and Context (b = 0.555, p < 0.001). The last interaction, that is of most interest, is a reflection of the fact that the performance level for the two different deviant stimuli [ɪpapu] and [ɛpapu] strongly differs depending on whether the "papu" context comes from a speaker with a low or a high F_1 range. In a low F_1 context participants found it harder to detect an [ɛpapu] deviant than an [ɪpapu] deviant whereas in a high F_1 context this effect was in the opposite direction.

A three-way interaction was found between the factors Stepsize, Deviant and Context (b = 0.181, p = 0.024), which reflects the fact that the critical interaction between Deviant and Context is bigger for the larger step size. A three-way interaction was found between Stepsize, Deviant and Block (b = 0.087, p = 0.040). This small effect reflects that, for [ɪpapu], performance for small step sizes increased as the experiment progressed whereas performance on large step sizes shows a tendency to decrease. This difference between small and large step sizes is not present for the discrimination of [ɛpapu]. A three-way interaction was found between Stepsize, Context and Block (b = -0.159, p< 0.001), a reflection of the fact that the two-way interaction between Stepsize and Block is stronger for the High F_1 condition. A three-way interaction was also found between Deviant, Context and Block (b = 0.085, p = 0.013), which shows that only [ɛpapu] targets in the high F_1 condition are increasingly better discriminated as the experiment progresses, whereas the targets in the other conditions do not show this pattern.

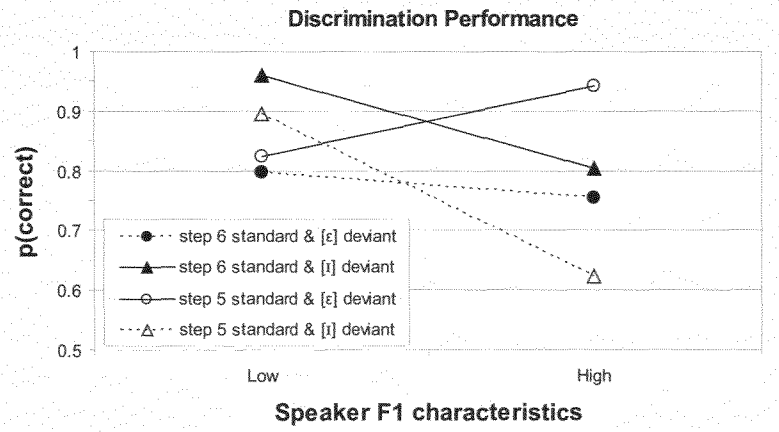

Figure 2. Experiment 2: Mean probability of a correct discrimination response to pairs of stimuli in the 4I-oddity design. Listeners performed a discrimination task in both a Low- F_1 and a High- F_1 speaker condition (defined by the F_1 value in the [papu] part). Deviant stimuli consisted of either [ɪpapu] ([ɪ]-deviant) or [ɛpapu] ([ɛ]-deviant), while the standards consisted either of step 6 (filled symbols) or step 5 (open symbols). Because of these combinations, pairs differed by either a small (4 steps, or 80 Hz: dotted lines) or large (5 steps, or 100 Hz: solid lines) frequency step.

Discussion

In a 4I-oddity vowel-discrimination experiment, it was found that listeners' discrimination performance was influenced by the speaker context in which vowels were presented. When participants listened to vowels in a low- F_1 speaker context, participants found it harder to detect an [ɛpapu] deviant than an [ɪpapu] deviant (with an ambiguous vowel as a standard) whereas in a high F_1 context this effect was reversed. This suggests that extrinsic speaker context not only adjusts category boundaries but also changes pre-categorical percepts.

One surprising aspect about the data from Experiment 2 is the asymmetry (see Figure 2) in the discriminability of the small step sizes (80 Hz) between the high and low F_1 condition. In the analysis of the data this asymmetry is expressed in the interaction between Stepsize and Context. We do not have a satisfactory explanation for this asymmetry but it should be noted that the effect is relatively small ($b = 0.181$, $p = 0.024$), especially when compared to the size of the critical Deviant by Context interaction ($b = 0.555$, $p < 0.001$).

The finding of the Deviant by Context interaction supports the idea that vowel normalization takes place at a pre-categorical level. An alternative possibility, however, is that despite the nature of the 4I-oddity vowel discrimination task that was used (encouraging listeners to focus on auditory properties, see e.g., Gerrits and Schouten, 2004), the listeners in Experiment 2 still focused mainly on the phonemic labels. If this were the case, vowel normalization in discrimination could still be the result of compensation through a shift in vowel categories. A third experiment was set up to test this possibility. In this experiment it was tested whether listeners indeed focus mainly on a categorical phoneme level in a design similar to the one in Experiment 2. Instead of only testing the discrimination of endpoints from ambiguous steps, we now tested how well participants were able to discriminate three-step differences over the whole continuum. That is, the first step was paired with the fourth step, the second step was paired with the fifth step and so on. In order to clarify the interpretation of the shape of the discrimination functions in Experiment 3 with three-step discrimination, it was first established what the shape of the discrimination functions would look like if participants focus on categorical representations. The next section reports on the simulated discrimination function that was calculated based on the categorization data from Experiment 1. This simulated experiment reflected discrimination behavior that would have been found if participants had relied on only categorical representations of the stimuli. This "simulated" categorization dataset was used to determine how a discrimination peak would be expected to express itself over the continuum.

Simulated Discrimination

Results

The upper panel of Figure 3 shows average predicted discrimination results. The data points were computed by aggregating over the response data from Experiment 1 for each of the factors Step, Context and Subject. Next, the predicted discrimination accuracy was calculated with the equation: accuracy = 0.5 + 0.5(p(step A) - p(step B))^2, where p(step A) and p(step B) are the proportions of [ɪpapu] responses, respectively, for step A and step B of a pair (equation adapted from Pollack and Pisoni, 1971). The analyses included the numeric factors for the main effects of Step (Levels: -3 (1 paired with 4) to 3 (7 paired with 10), in steps of 1), Context (Levels -1 = Low F_1 and 1 = High F_1) and Step-squared (that was used to test for the

presence of a discrimination peak in the middle of the continuum) and interaction terms for interactions between Step by Context and Step^2 by Context.

The optimal model had main effects for the Intercept ($b = 0.637$, $p < 0.001$), indicating predicted discrimination that is higher than chance; and Step^2 ($b = -0.012$, $p < 0.001$), that shows that the predicted discrimination is higher in the middle of the continuum than towards the two endpoints. An interaction was found between the factors Context and Step ($b = 0.013$, $p = 0.002$). This indicates that the predicted discrimination performance is higher towards the [ɛ] endpoint for the High F_1 condition and higher towards the [ɪ] endpoint for the Low F_1 condition. This is a reflection of the fact that the discrimination peak is expected to shift in different directions for the two context conditions.

Discussion

The predicted discrimination functions showed two general patterns. The first was that the predicted data reflected significant discrimination peaks. These peaks were located at the category boundaries for the two different speaker conditions. The second pattern was that the discrimination functions had a different slope for the two different speaker conditions (a result of the fact that the peaks were located at different vowel pairs).

In Experiment 3, discrimination performance was tested on the same range of sound-pairs over the complete [ɪ] to [ɛ] continuum, but then with actual participants performing the discrimination task. A categorical discrimination strategy would be revealed by significant effects for both of the patterns found in the predicted data: Discrimination peaks at the category boundaries and different slopes of the discriminability function along the continuum. A pre-categorical discrimination strategy, however, would result in only the second pattern: The different context conditions would lead to differences in discriminability gradually developing over the range of vowel stimuli, expressed by a difference in the slope of the discrimination functions for the two context conditions. Discrimination should then be high at the [ɪpapu] end of the continuum for the Low- F_1 context condition and high at the [ɛpapu] end of the continuum for the High- F_1 condition.

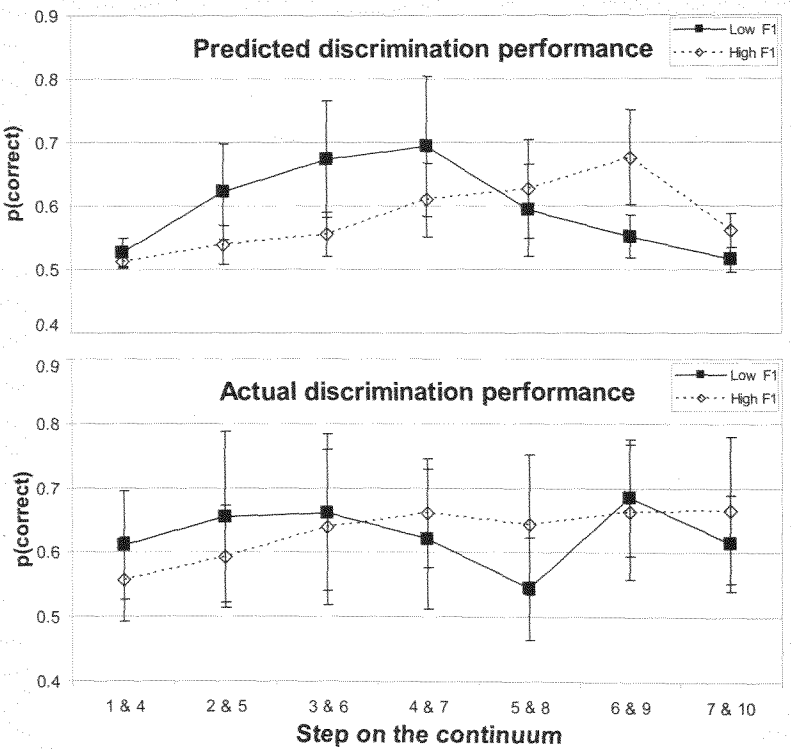

Figure 3. Upper panel: Predicted mean probabilities of a correct response to pairs of stimuli in a 4I-oddity design. Predictions were based on the categorization data collected for Experiment 1. Discriminability is predicted for "Vpapu" pairs for which the F_1 of the initial vowel sounds differ by 3 steps (or 60 Hz), sliding across the continuum from 1 to 10 (e.g., pairs of steps 1 & 4, 2 & 5, etcetera). This resulted in predicted values for 7 steps. Discrimination performance is predicted for both the Low F_1 and the High- F_1 speaker condition (defined by the F_1 value in the [papu] part). Error bars reflect predicted confidence intervals per step. Lower panel: Experiment 3: Actual mean probabilities of correct responses. Error bars reflect confidence intervals per step.

Experiment 3

Method

Materials and procedure.

The stimuli that were created for Experiment 1 were used for the creation of the 4I-oddity pairs. Members of a pair were concatenated to create sets of four stimuli with an interstimulus interval of 500 ms. These stimuli consisted of three standard items and one deviant item that was either in second or third position. Combinations consisted of pairs with an F_1 frequency difference of 60 Hz (e.g., 1 paired with 4, 2 paired with 5 etcetera up to 7 paired with 10), resulting in 7 pairs. Two versions of these were created such that both items could occur as either standard or deviant (e.g., step 10 as standard and 7 as deviant and step 7 as standard and step 10 as deviant). This resulted in 7 (step on the continuum) * 2 (which item of the pair is the deviant) * 2 (context speaker F_1 characteristics) = 28 types of discrimination stimuli, with two possible deviant positions (in second or third position). Stimuli from the two different context conditions were presented to two different groups of participants.

Stimuli were presented in blocks of sixteen stimuli with a balanced number of trials where the deviant was in second or third position (presented in a randomized order). Two versions of these, with the same step from the continuum as either the standard or the deviant, were presented in successive blocks. The different pairs from the 7-step continuum were presented as 7 overall blocks in a randomized order. After 112 stimuli (halfway through the experiment) participants were allowed a self-paced pause. Response options were always on the screen (printed "2"and "3"). Participants were allowed to respond throughout the duration of the trial and responded by pressing one of two buttons on a button box. Participants received visual feedback (printed "Correct" (correct) or "Fout"(incorrect) on the computer screen) immediately after their responses.

Participants.

Twenty-four participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. These were selected according to the same criteria as those for the first experiment. They received a monetary reward for their participation

Results

Responses faster than 100 ms after the onset of the second item (the first possible deviant) were excluded (99.9% of the data were kept). Figure 3b shows the average discrimination results. Analyses included the numeric factors Step (Levels: -3 (1 paired with 4) to 3 (7 paired with 10), in steps of 1), Step^2, Block (Levels -3 = first block to 3 = last block in steps of 1) and Context (Levels -1 = Low F_1 and 1 = High F_1). All interactions were included except for those that contained the Step by Step^2 interaction. The optimal model for the data had significant main effects for the Intercept ($b = 0.5567$, $p < 0.001$), reflecting the fact that participants discriminated above chance level; and the factor Block ($b = 0.0686$, $p< 0.001$), that reflects an increasing number of correctly discriminated stimuli as the experiment progressed. An interaction was found between the factors Step and Context ($b = 0.0480$, $p < 0.001$). This critical effect shows that the degree of discriminability across the steps of the continuum is different for the two speaker context conditions.

If the factor Step^2 was added to this model it had a non significant effect ($b = -0.0074$, $p = 0.3715$). This shows that participants' actual discrimination responses, unlike those predicted from categorization data, did not result in a discrimination peak.

Discussion

Discussion

Experiment 3 was set up to investigate whether participants relied on phonemic labels while performing the 4I-oddity task. The discrimination behavior that was predicted from the categorization data of Experiment 1 showed that a categorical discrimination strategy would have resulted in significant discrimination peaks at different steps on the continuum in each context. The actual discrimination performance that was tested in Experiment 3 did not result in any significant discrimination peaks. This shows that the discrimination data in Experiment 3 (and by extension Experiment 2) largely reflect participants' perception of the auditory properties of the stimuli.

A second motivation for Experiment 3 was to see how speaker context influences perception across the continuum. The data (see Figure 3) show that the High and Low- F_1 speaker contexts influence the discriminability of vowel pairs differently across the continuum. For vowels presented in a High- F_1 context it was easier to discriminate sounds on the [ɛ] (towards step 10) side of the continuum

whereas participants generally found it easier to discriminate sounds towards the [ɪ] end of the continuum in a Low- F_1 context. The shift in discriminability seems to change in a relatively smooth manner. There is one data point that does not follow the general trend however: step 6 & 9, in the Low F_1 condition. It is unclear what the cause of the local increase in discriminability is. Importantly, however, it does not reflect a category boundary effect: As Figure 3 shows, an increase in discriminability due to a category boundary in this speaker condition would be expected towards the other end of the continuum.

General Discussion

The series of experiments reported here investigated the cognitive locus of extrinsic vowel normalization. The first experiment showed that listeners interpret an [ɪ] to [ɛ] continuum differently depending on the F_1 range of a speaker in a context stimulus. This is a replication of earlier findings that have shown that perceived vowel identity can be influenced by a preceding sentential context (Ladefoged & Broadbent, 1957; Sjerps, et al., 2011; van Bergem, et al., 1988; Watkins, 1991; Watkins & Makin, 1994, 1996). In the experiments reported here, however, an effect was established using a speaker context that was only two syllables long (/papu/). This context was spliced after the target vowel (targets were of the general structure: /Vpapu/). There was always a 500 ms interval between a preceding context and a target vowel. This approach prohibited potential peripheral auditory influences (Summerfield, et al., 1984; Wilson, 1970). Peripheral auditory influences show an effect in the same direction as normalization effects that were under investigation and are as such indistinguishable from those effects. Peripheral auditory effects strongly diminish after a longer ISI. This allowed us to focus on central compensation processes (Watkins, 1991).

In the second experiment these stimuli were used in a 4I-oddity discrimination experiment. The phoneme category shift proposal assumes that vowel normalization is the result of a shift of category boundaries, taking place at a categorical level of processing. As such, it predicted that Experiment 2 would not lead to vowel normalization effects. The auditory proposal (Watkins, 1991; Watkins & Makin, 1994, 1996), however, assumes that compensation effects arise at auditory levels. This proposal thus predicts that vowel normalization effects will express themselves not only in categorization but also in an auditory-focused task like the 4I-oddity task. The results confirmed the predictions made by an auditory proposal: discrimination

performance was dependent on speaker context. In the context of a speaker with a high F_1, listeners found it more difficult to discriminate between an [ɪ] and an ambiguous sound than between an [ɛ] and an ambiguous sound. In the context of a low F_1-speaker this pattern was reversed.

The third experiment confirmed that normalization effects in 4I-oddity discrimination experiments are pre-categorical in nature. The discriminability of vowel pairs across the [ɪ] to [ɛ] continuum was again influenced by the speaker context but, critically, did not show a discrimination peak at the category boundaries. A discrimination peak in the discrimination function would have meant that the 4I-oddity design that was used here still encouraged listeners to rely on their category labels. This would have made it impossible to attribute the shifts in discriminability to the pre-categorical processing level. Simulated discrimination functions based on the categorization data of Experiment 1 showed how a categorical strategy would have been expressed in discrimination scores. Unlike the predicted discrimination functions based on Experiment 1, the discrimination functions in Experiment 3 did not show discrimination peaks at the (contextually-conditioned) category boundaries[2]. This shows that listeners indeed focused on pre-categorical properties of the vowel stimuli that were used here.

The simplest account of the normalization effects observed with the same stimuli across all three current experiments, therefore, is that the effects all reflect, at least in part, a pre-categorical process. Note, however, that this investigation, does not determine whether these pre-categorical representations already contain some form of abstraction (e.g., abstractions over auditory cues). Nevertheless, this study does show that normalization can take place at a level that precedes the level at which phonemic

[2] Repp et al. (1979) suggested that the correspondence between identification and discrimination performance for vowel stimuli needs to be assessed in the same format, rather than performing an identification task with one stimulus per trial and a discrimination task with two stimuli in a row (or four stimuli in our case). They based their suggestion on the observation of contrast effects when stimuli are identified in the format of an AX discrimination task. They suggest that the identification of the two vowels is strongly influenced by their adjacent co-presence in a classical AX task. This is unlikely to be an issue in the current case because the context preceding the to-be-identified or to-be-discriminated vowels was identical in both tasks (namely the papu part of the preceding stimulus). Repp et al. (1979) also showed that such intervening materials abolish any contrast effect on the target vowels (see their Table 4). The only difference then between the situation during identification and discrimination is the inter-stimulus interval, which is longer and dependent on the speed of the participant's response in identification. However, the ISI was set to a minimum of 500 ms in the discrimination task to prevent effects of auditory after-image and it stands to reason that additional time beyond these 500 ms would make little difference. As such, it is quite unlikely that the lack of correspondence between identification and discrimination in the present data is due to the different task settings.

judgments are made. The fact that a sound that is ambiguous between [ɛ] and [ɪ] is more often labeled as /ɪ/ in a low F_1 context and more often as /ɛ/ in a high F_1 context (see Figure 1) thus appears to involve a shift in perceptual space that takes place at a pre-categorical level. This provides support for the notion that an important component of extrinsic normalization is a process which takes place at an auditory level of representation (Sjerps, et al., 2011; Watkins, 1991; Watkins & Makin, 1994, 1996).

A purely auditory approach, however, is not able to explain the findings reported by Johnson et al. (1999) of non-auditory contextual influences on the categorization of vowel continua. They report vowel categorization shifts that were induced by asking participants to imagine listening to a male or a female speakers. In this case learned covariations between speaker gender and average F_1 influence subsequent categorization. The combination of the findings by Johnson et al. (1999) and the findings reported here suggest that vowel normalization effects do not arise from a single processing level. Normalization effects could thus have both pre-categorical and higher level cognitive components. The effects observed in Experiments 2 and 3, due to the nature of the 4I-oddity task as revealed especially by the absence of a discrimination peak at the category boundary in Experiment 3, can mainly be attributed to pre-categorical processes. The effect observed in Experiment 1 might have a cognitive component though. The important point, however, is that a major component of the effect in a categorization task like Experiment 1 is likely to be the result of a restructuring of perceptual space at a pre-categorical level of processing. If this low-level process is at work in the 4I-oddity task, there is every reason to suppose that it is also at work in a categorization task.

To our knowledge this is the first study that tests influences of extrinsic speaker vocal tract characteristics on the discriminability of vowels. Similar approaches have been taken in the closely related fields of compensation for coarticulation (Stephens & Holt, 2003) and compensation for assimilation (Mitterer, et al., 2006) with similar results. Mitterer et al. (2006) report on a Hungarian assimilation rule: The final [l] in the syllable [bol] is produced as [l] in "bolnal":[bɔlna:l], but as [r] in "bolrol":[bɔrro:l]. Hungarian listeners compensate for this assimilation rule because they perceive more sounds on a [l] to [r] continuum as [l] in the assimilation-viable word "bolrol" (compared to the word "bolnal"). This

does not depend on language background because in their study Mitterer et al. (2006) found that Dutch listeners showed the same compensation effect (Dutch does not have this rule, nor were the listeners familiar with Hungarian). Interestingly, the same effect was found with discrimination: Listeners, both Dutch and Hungarian, found it harder to distinguish between [bol] and [bor] in the assimilation-viable "...rol" context than in the assimilation-unviable "...nal" context. Additional findings showed that non-speech contexts (non-speech versions of the [rol] and [nal] syllables) induced similar effects. Mitterer et al. (2006) argue that their findings support the idea that these kinds of compensation effects take place at an auditory processing level.

In a similar vein, Stephens and Holt (2003) tested whether compensation for coarticulation generalizes to the influence of speech contexts on non-speech targets. They elaborated on the finding by Mann (1980) that an ambiguous CV syllable halfway between [ga] and [da] is more often identified as /ga/ (that has a low F_3) when preceded by [al] (high F_3), but more often identified as /da/ (high F_3) when preceded by [ar] (low F_3). Stephens and Holt (2003) found that the perception of both speech and non-speech versions of the [ga] - [da] continuum could be influenced by the [al] vs. [ar] preceding speech context. Because non-speech targets cannot easily be categorized by participants they used a discrimination design for which category labels are not necessary. Stephens and Holt (2003) showed that it was more difficult to discriminate steps on the continuum when the preceding syllables influenced the targets such that the F_3 values were perceptually brought closer to each other than when the context acted to increase the perceived difference in F_3. This effect was found for both speech and non-speech targets. These findings suggest that this type of compensation for coarticulation also has an auditory basis. These reports, that all investigated the contrastive nature of speech perception, therefore all came to the same conclusion: Compensation/normalization processes can largely be attributed to auditory processes because the effects of these processes are visible in discrimination tasks.

Although the experiments reported here were not set up to specifically test this issue, an additional suggestion based on the discrimination results in Experiments 2 and 3 is that they do not seem to be the result of a shift of the complete perceptual space. Rather, the restructuring of perceptual space seems local. That is, the influence of the F_1 manipulation is restricted to frequency regions that are close to the average F_1 in the context. A shift of perceptual space over the complete frequency range

should have led to no differences in discriminability over the two context conditions. For example, consider Experiment 3: If vowel normalization were to result in an auditory transformation that shifted all frequencies up by 100 Hz then the differences between the F_1 of the stimulus pairs should remain to be 60 Hz. Discriminability across the continua should then not differ for the two context conditions.

To conclude, this paper reports on extrinsic normalization effects like those first reported by Ladefoged and Broadbent (1957). These effects were established with a paradigm that is to our knowledge novel in research on extrinsic vowel normalization: one based on a comparison between categorization and discrimination tasks. The use of a discrimination task is valuable in the investigation of extrinsic vowel normalization. It is a task that does not rely on the use of phoneme categories and reduces possible higher-level influences. Such influences can hide or exaggerate normalization effects by introducing biases that cause shifts in categorization functions. Here, the comparison between categorization and discrimination behavior established that extrinsic vowel normalization influences perception at an early, pre-categorical level of processing.

Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics

---

Chapter 6

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (in press). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia.*

Abstract

This study used an active multiple-deviant oddball design to investigate the time-course of normalization processes that help listeners deal with between-speaker variability. Electroencephalograms were recorded while Dutch listeners heard sequences of non-words (standards and occasional deviants). Deviants were [ɪpapu] or [ɛpapu], and the standard was [ˈɛpapu], where [ˈɛ] was a vowel that was ambiguous between [ɛ] and [ɪ]. These sequences were presented in two conditions, which differed with respect to the vocal-tract characteristics (i.e., the average 1st formant frequency) of the [papu] part, but not of the initial vowels [ɪ], [ɛ] or [ˈɛ] (these vowels were thus identical across conditions). Listeners more often detected a shift from [ˈɛpapu] to [ɛpapu] than from [ˈɛpapu] to [ɪpapu] in the high F_1 context condition; the reverse was true in the low F_1 context condition. This shows that listeners' perception of vowels differs depending on the speaker's vocal-tract characteristics, as revealed in the speech surrounding those vowels. Cortical electrophysiological responses reflected this normalization process as early as about 120 ms after vowel onset, which suggests that shifts in perception precede influences due to conscious biases or decision strategies. Listeners' abilities to normalize for speaker-vocal-tract properties are for an important part the result of a process that influences representations of speech sounds early in the speech processing stream.

Introduction

In everyday life, we listen to the speech of many individuals. The current paper investigates a perceptual compensation process that helps listeners to understand speech sounds spoken by different talkers. Individuals have different vocal-tract characteristics, caused by influences such as talker sex, talker size (or vocal-tract length), speaking style, and dialect. This variance appears to challenge speech comprehension because vocal tracts can differ on the same acoustic dimensions that allow listeners to discriminate between different speech-sound categories. We ask here how early in the speech perception process listener's representations of speech sounds are changed in order to compensate for talkers' vocal-tract characteristics.

Vowels are discriminated mainly on the basis of acoustic properties that are referred to as formants. Formants are bands of increased intensity in the spectral makeup of speech sounds. For example, in English the main difference between the words "bit" and "bet" (phonemically transcribed as /bɪt/ vs. /bɛt/) lies in the frequency of the first formant (F_1). The average F_1 value for /ɛ/ is around 731 Hz while the average F_1 of /ɪ/ lies around 483 Hz, for vowels recorded from female American English speakers (F_2 value for /ɛ/: 2058 ; F_2 of /ɪ/: 2365). For male speakers the average F_1 value for /ɛ/ is around 580 Hz while the average F_1 of /ɪ/ lies around 427 Hz (F_2 value for /ɛ/: 1799 ; F_2 of /ɪ/: 2034) (Hillenbrand, et al., 1995). However, these averages do not tell the complete story. There is a large degree of overlap among different vowel categories (Hillenbrand, et al., 1995; Joos, 1948). Single instances of two different vowels, spoken by two different speakers, can have very similar absolute formant values. This is not restricted to English; Dutch, the target language of the current paper, shows similar overlap in vowel categories (Adank, van Hout, et al., 2004; Van Nierop, Pols, & Plomp, 1973). This is especially the case when comparing speakers of different sex or age. Such variance therefore causes multiple signal-to-category mappings for a single spoken speech sound. In other words, a single sound can often be interpreted as either of two different phonemes, so that listeners may be confused whether the intended word was *bit* or *bet*.

It has been argued that listeners compensate for vocal-tract characteristics in a number of different ways (Johnson, 2005; Nearey, 1989). An important contribution may be made by a mechanism that compensates for speaker characteristics by taking into account the vocal-tract characteristics of the speaker as revealed in a preceding context (Ladefoged & Broadbent, 1957). Ladefoged and Broadbent (1957) found that

listeners interpret a vowel that is acoustically halfway between an [ɪ] and an [ɛ] more often as /ɪ/ (which has a low F_1) when it is preceded by a sentence with a relatively high F_1, while the same sound is more often interpreted as /ɛ/ (which has a high F_1) when preceded by a precursor sentence with a relatively low F_1. This contrastive process therefore effectively normalizes perception for the F_1 range of a speaker and reduces potential overlap of vowel categories across speakers. Although F_1 is not the only cue to differences in vocal-tract characteristics, Ladefoged and Broadbent (1957) have shown that listeners can use F_1 characteristics to map speech sounds onto the correct phonemes.

Watkins (1991) and Watkins and Makin (1994, 1996) have argued that the bulk of this effect can be explained by a mechanism that compensates for the average spectral makeup of a precursor, whether the shape of this spectral makeup is due to vocal-tract characteristics or something else (such as room acoustics). The suggestion is that these context effects are the result of a mechanism that focuses on how a given stimulus is different from the preceding context ("Sensitivity to *change* ", cf. Kluender, Coady, & Kiefte, 2003). This mechanism is assumed to be a general perceptual mechanism that is not specific to speech perception. In line with this claim, contrast effects similar to vowel normalization also occur for musical timbre perception (Stilp, et al., 2010). Additional evidence stems from the finding that a precursor spoken by a female talker can influence the perception of a subsequent target sound that was produced by a male talker (Watkins, 1991), and that speech sound categorization can also be influenced by non-speech precursors (Holt, 2006a). This kind of mechanism could therefore function as a means of enhancing contrast (Kluender, et al., 2003; Kluender & Kiefte, 2006) and displays a clear analogy to contrast effects with visual stimuli. A surface with a certain brightness will be perceived as darker when surrounded by a light surface, but as lighter when surrounded by a dark surface. Furthermore, effects of preceding context on speech sound categorization have also been observed with Japanese quail (Lotto, et al., 1997). This again suggests that influences of context are in part the result of a relatively general perceptual mechanism (see Holt, 2006, for an overview of context dependent effects on categorization), but it is not clear that a single general-purpose mechanism is sufficient to explain all the results on vowel normalization. Sjerps, Mitterer and McQueen (2011) argue, for example, that vowel normalization may primarily reflect a compensation mechanism that is based on the Long-Term Average

Spectrum (LTAS) of the auditory input, but one that only operates if the input has spectro-temporal characteristics that are similar to those of speech. The primary compensation mechanism for vowel normalization thus appears to be general-purpose (based on contrast), but one which operates under some spectro-temporal constraints.

Little is known, however, about when in the processing stream this compensation mechanism has its influence. Does normalization influence low-level representations or does it influence higher-level cognitive processes? Clearly, the assumption of a general perceptual mechanism that focuses on change is more compatible with the assumption of an early locus. The present study therefore examines the temporal locus of compensation for speaker vocal-tract characteristics by tracking its neurophysiological correlates during the perception of vowels. This is a novel approach in the investigation of the extrinsic normalization of vowels.

In order to establish whether normalization influences representations early or late in the stream of processing, four different time windows were investigated that can be considered to reflect subsequent stages in the processing stream. These were the P1, the N1, the N2 and the P3 time windows. Previous neurophysiological investigations with speech stimuli have suggested different functional interpretations for the processes/representations underlying these different waveform components. The earliest long-latency brain waves (P1 and N1 or their magnetic counterparts P1m, N1m) peaking at about 50 and 100 ms after stimulus onset respectively, seem to reflect early cortical information processing (Diesch, Eulitz, Hampson, & Ross, 1996; Makela, Alku, & Tiitinen, 2003; Obleser, Eulitz, & Lahiri, 2004). While P1 has been argued to reflect basic auditory feature extraction, N1 seems to reflect a subsequent level closer to a more abstract phonological representational stage (Tavabi, Obleser, Dobel, & Pantev, 2007). Roberts et al. (2004) argue that the N1 response represents some form of abstract processing. Cortical responses that were recorded when participants listened to sounds on a vowel continuum from [u] to [a] reflected clustering of N1 peak latencies around the regions of the continuum identified as either [u] or [a] (and less clustering around the ambiguous region). Roberts et al. (2004) also find, however, that when acoustic aspects of a single stimulus are held constant while the percept is changed through a response bias induced by preceding trials, the dominant N1 latency effect is related to the physical properties of the stimulus, and not to the eventual decision. This indicates that higher-level processes like response biases do not influence N1. Furthermore, while the N1 might reflect

some abstract properties, Näätänen and Winkler (1999) argue that it does not reflect a completely abstract level of processing as it does not directly reflect the consciously perceived event. Bien, Lagemann, Dobel, and Zwitserlood (2009) have also shown that there is a difference between the N1 response and conscious decisions about the signal. Furthermore, in contrast to Roberts et al. (2004), Toscano, McMurray, Dennhardt, and Luck (2010) have shown that, for the perception of the voiced versus voiceless stop consonant distinction (as in beach versus peach), there is no relation between the N1 amplitude and the categorical status of a phoneme. They found that there was a linear relation between the N1 amplitude and the step on the voiced-voiceless continuum. Thus, while it is not yet clear whether the N1 can reflect abstract aspects of speech sounds, previous results do show that the N1 reflects processes that are not influenced by response-bias or by the consciously perceived qualities of the stimulus. This shows that these processes take place early in the processing of speech information. Roberts et al. (2004) thus argue that the ultimate perception of speech sounds depends on the coding of stimulus properties that takes place during the N1 time window.

Later time windows such as the N2/MMN time window (200 - 300 ms after stimulus onset) do seem to reflect abstract levels of processing (Näätänen et al., 1997; Winkler et al., 1999). MMN responses, for example, are larger to deviants that are linguistically relevant for the listener (Näätänen, et al., 1997; Winkler, et al., 1999). Moreover, in a study measuring both N1 and MMN, Sharma and Dorman (2000) found a dissociation between measures of N1 and MMN. Both Hindi and English listeners showed similar, direct dependencies of N1 latency on a Voice Onset Time (VOT) continuum that is only relevant for listeners of Hindi (-90 to 0 ms). However, only Hindi listeners elicited a MMN effect with these stimuli. This shows that N1 and MMN reflect subsequent stages in the processing hierarchy and only the MMN response is dependent on linguistic exposure. Finally, the P3 response in response-active oddball designs (300 - 600 ms after stimulus onset) has been associated with the evaluation of deviant events with relation to subsequent behavioral action (Friedman, Cycowicz, & Gaeta, 2001). The P3 is thus likely to also reflect higher-level cognitive processes, although it is not necessarily insensitive to gradedness within speech categories (Toscano, et al., 2010).

Our aim here was to investigate whether compensation for speaker vocal-tract characteristics is a process that influences representations of speech sounds at a

relatively early stage of processing (i.e., during P1 and/or N1 time windows) or at a relatively late stage of processing (i.e., during N2 and/or P3 time windows). We investigated the influence of vocal-tract characteristics on vowel perception by presenting participants with target vowels in contexts that simulate speakers with different vocal-tract characteristics. Previous findings have shown that manipulated context sentences can change the perception of subsequently presented vowels, indicated by a shift in the categorization functions for these vowels (Kiefte & Kluender, 2008; Ladefoged & Broadbent, 1957; Mitterer, 2006a; Sjerps, et al., 2011; Watkins, 1991; Watkins & Makin, 1994, 1996).

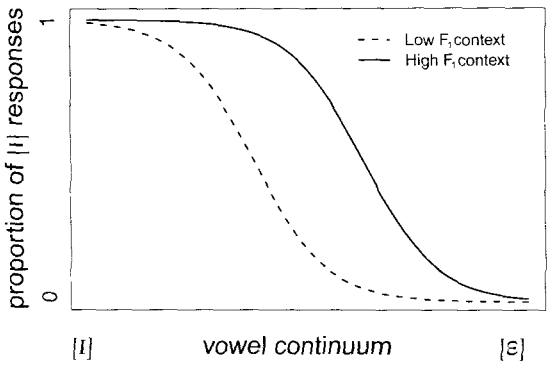

Figure 1 Schematized response functions: mean proportions of an [ɪ] response to a range of target sounds from [ɪ] to [ɛ], presented after a precursor with an increased F_1 (High-F_1 context) or decreased F_1 (Low-F_1 context) level.

Consider Figure 1. It displays two mock-up categorization functions for sounds from a vowel continuum ranging from [ɪ] to [ɛ], that represent the sort of shift in categorization function that has been found. The dotted line represents the categorization of vowels that have been presented after a precursor with a generally low F_1, while the solid line represents categorization of the same vowel tokens, but then presented after a precursor sentence with a high F_1. It can be observed that more sounds on the continuum are categorized as /ɪ/ (which itself has a low F_1) in the context of a precursor with a high F_1 and more often as /ɛ/ (which has a high F_1) in the context of a low F_1 precursor. The current research attempts to investigate at what

level of processing the representation of speech sounds is influenced by the mechanism that leads to this shift in perception.

In the present experiment, target non-words were presented in a response-active mismatch detection design, such that listeners heard a repeating (standard) non-word that was replaced by two different (deviant) non-words on 20% of trials. The standard consisted of a non-word in which the initial vowel was manipulated to sound halfway between [ɪ] and [ɛ], from now on indicated by [ˈɛ] (the transcription of the ambiguous sound as [ˈɛ] should make clear that this sound does not represent an actual Dutch phoneme category). The deviant non-words started with a vowel that was an unambiguous instance of /ɪ/ or /ɛ/. The following two syllables in each non-word (/papu/) were manipulated to have a high F_1 or a low F_1 so as to induce normalization effects in different experimental blocks. The bisyllable /papu/ contains two point vowels that provide the range of a speaker's F_1. The induced change in perception through normalization should make it harder for participants to detect a change from the ambiguous vowel [ˈɛ] to [ɪ] than to [ɛ] in the high F_1 context, whereas listeners should find it harder to detect a change from [ˈɛ] to [ɛ] than to [ɪ] in the Low F_1 context.

Listeners thus heard the nonsense words [ˈɛpapu] (as the standard stimulus), and [ɪpapu] and [ɛpapu] (as the deviant stimuli). In this setup the 2nd and 3rd syllables of stimulus x provided the preceding context for the next stimulus, x+1. This approach was chosen to be able to create an interstimulus interval (ISI) between the context ([papu]) and the subsequent target-vowels of 750 ms (i.e., [ˈɛpapu] - 750 ms - [ɛpapu] - 750 ms - [ˈɛpapu] etcetera). The contextual influence of the [papu] syllables might extend to subsequent trials (i.e., the perception of a target vowel is influenced not only by the just-preceding context), but because of the blocked presentation this influence was in the same direction within a block. The large ISI between a target vowel and its immediately preceding context is important because small ISI could lead to contextual influences that are a result of peripheral auditory influences such as the negative auditory after-image (Summerfield, et al., 1984; Watkins, 1991; Wilson, 1970). Such peripheral influences can cause a compensation effect in the same direction as the more central compensation effect under investigation here and could thus obscure its effects. The contexts thus followed directly after the target vowels. This is not a problem for the interpretation of the EEG waveform, however, as the following context started 250 ms after the onset of the vowel, which leaves enough

time for any early cortical signatures in response to the critical vowels to appear before any effect of following context (at least those in the P1, N1, and N2 time windows). The early components induced by the following [papu] coincide with the P3 effect induced by the target vowel. This is not a problem, however, as the P3 response is larger in amplitude than the earlier cortical responses that could be induced by the following context. It should be noted that the strength of normalization effects might decrease over repetitions (Broadbent & Ladefoged, 1960). This decrease, however, has been argued to be stronger when different context conditions are presented in a mixed fashion instead of the blocked approach that was taken here (Sjerps, et al., 2011).

An additional control condition was run that had the vowel [ɔ] as the initial vowel on the standard items (i.e., [ɔpapu]), but had the same deviants as in the experimental condition ([ɪpapu] and [ɛpapu]). In this control condition the [papu] part had a neutral F_1 contour that was halfway between that of the high and low F_1 conditions. This control condition was used to test whether our design was capable of producing a clear standard-deviant mismatch effect in the cortical signatures, and when and where on the cortical topography these mismatch effects would express themselves.

The control data were analyzed by comparing the size and distribution of the effect of deviant ([ɪpapu] and [ɛpapu]) versus standard ([ɔpapu]) in the four time windows. For the experimental (i.e., non-control) stimuli, an initial analysis compared ERPs between the two standard stimuli ([ˈɛpapu] in both the High F_1 and the Low F_1 condition) versus the deviants ([ɪpapu] and [ɛpapu] in both the High F_1 and the Low F_1 condition). This comparison was made to see whether and when the small auditory differences that we used in the experimental condition were able to elicit different cortical responses to deviants (note that in both sets of data the deviant vowels were the same, only the standards differed). In the final and critical analysis we tested at what point in the stages of cortical processing of speech the influence of the contexts' F_1-properties on the detectability of a vowel change was reflected. This effect was tested by looking for an interaction between the F_1 condition and the identity of the deviant vowel, with the size of the difference response (in voltage) as the dependent variable. Note that the analysis of this critical interaction focuses on the processing of the deviants and not on the processing of the standard, despite the fact that our design hinges on the fact that the perception of the standard is changed across blocks.

Normalization processes change the perceived quality of the standard and thus also the mental traces of the standard. For the critical analysis, we measured the relative strength of the cortical signature of the mental comparison of a deviant vowel to those traces of the standard. Traditionally, designs with this oddball paradigm focus on the difference wave between standard and deviant (cf. Näätänen & Winkler, 1999). As we were interested primarily in the interaction between deviant identity and the contexts' F_1 properties, a comparison of the deviants themselves suffices here.

In the present study, an early influence should thus be reflected in early time windows (i.e., within the first 160 ms after vowel onset) such as those related to the P1 and/or the N1, whereas a later influence should only be able to affect cortical signatures later than about 200 ms, a time window which is related to the N2/MMN or the P3. To exemplify the expected results, imagine the analysis in the P3 time window. We expected that easy detectability of deviants would lead to a stronger positivity. The [ɪ] deviant should be easier to detect (and thus result in a larger positivity) in the low F_1 condition than in the high F_1 condition. The difference wave for "[ɪ] in a low F_1 context" - "[ɪ] in a high F_1 context" should thus be positive. For [ɛ], this pattern should be reversed, and the difference wave for "[ɛ] in a low F_1 context" - "[ɛ] in a high F_1 context" should be negative. This mirror-image pattern of results should not necessarily arise only in the P3 time window; in fact, it should be observed from the point in time where the normalization processes start to take place. The question we asked was when that would be.

Method

Participants

Twenty-four native speakers of Dutch from the Max Planck Institute for Psycholinguistics participant pool were tested. They received a monetary reward for their participation. None of the participants reported a hearing disorder, language impairment, or uncorrected visual impairment and all participants were right-handed.

Materials

All recordings were made by a female native speaker of Dutch. Acoustic processing of the stimuli was carried out using PRAAT software (Boersma & Weenink, 2005). The materials consisted of the three-syllable nonsense sequences "ipapu" (/ɪpapu/) and "epapu" (/ɛpapu/), which are meaningless in Dutch. To create a continuum ranging from [ɪpapu] to [ɛpapu], the first vowel of a recording of [ɛpapu] was excised. This token had an F_1 frequency of 734 Hz, measured over a 40 ms

window at a steady part in the middle of the vowel. A continuum was created by using a Linear Predictive Coding (LPC) procedure to generate a source and filter model of the vowel. The frequency of F_1 was decreased in the filter model, which was then recombined with the source model. Three target steps were selected ranging from [ɛ] to [ɪ]. F_1 values ranged from 724 Hz ([ɛpapu]), through 634 Hz (the ambiguous item [$^{I}_{\varepsilon}$papu]), to 544 Hz ([ɪpapu]). The manipulated vowels were then low pass filtered at 1000 Hz and combined with the higher frequencies of the original vowel (1000-5000 Hz), to make them as naturally sounding as possible while retaining the perceived /ɛ/ versus /ɪ/ identity for the two endpoints. The created vowels were adjusted so that their amplitude envelope and overall amplitude matched that of the original sound. The top left panel of Figure 2 displays the spectra of the three vowels. The bottom left panel displays the difference between the LTAS of the endpoints [ɛ] and [ɪ]. It shows that [ɛ] has considerably more energy in the region between ~ 500 and 1000 Hz than [ɪ].

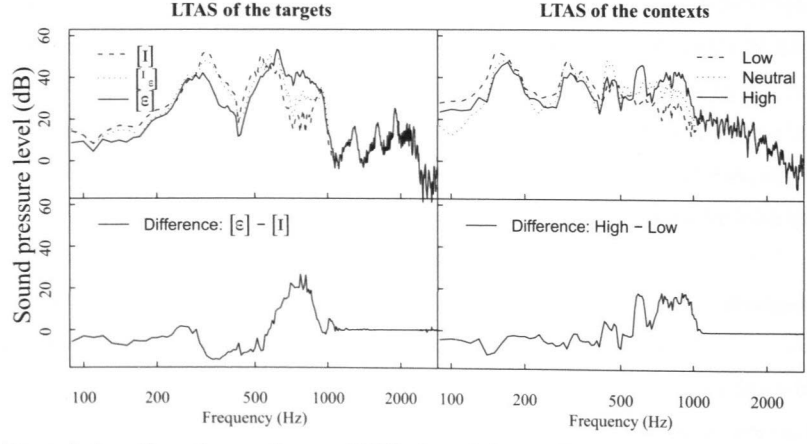

Figure 2. Long-Term Average Spectra (LTAS) plots of the stimuli. Upper left panel: LTAS for the deviant target vowels [ɪ] (dashed line) and [ɛ] (solid line), and the ambiguous standard vowel (dotted line). Left bottom panel: The difference spectrum for the deviant target vowels ([ɛ] - [ɪ]). Upper right panel: LTAS for the [papu] contexts that were manipulated to have a low (dashed line), a neutral (dotted line), or a high F_1 (solid line). Right bottom panel: the difference spectrum for the [papu] contexts (high - low). The x-axis is displayed on a logarithmic scale because this scale more accurately reflects the response properties of the auditory system.

To create the [ɔpapu] item for the control condition, an instance of [ɔ] was selected from a similar segmental context (an initial /ɔ/ followed by a /p/) and spoken by the same speaker. This sound was manipulated to have the same pitch, the same amplitude envelope and the same duration as the critical ambiguous vowel items.

The original recorded [papu] context had F_1 values of 730 Hz ([a]) and 410 Hz ([u]) measured over a 40 ms window at a steady part in the middle of the vowels. These two syllables were manipulated by the same F_1 manipulation procedure but then with either an increase of F_1 by 200 Hz, no increase, or a decrease of F_1 by 200 Hz to create the high F_1 context, the neutral context, and the low F_1 context respectively. The F_1 manipulation had some effect on the perceptual characteristics of the vowels in the [papu] sequence. Nevertheless, we will transcribe this sequence throughout as [papu], since this reflects the original utterance. The top right panel of Figure 2 displays the LTAS of the three contexts. The bottom right panel shows the difference between the LTAS of the high F_1 and the low F_1 contexts. The high F_1 context has more energy than the low F_1 context in the region between ~ 500 and 1000 Hz. This pattern is similar to that of the target vowels. A normalization mechanism that operates through compensation for LTAS could thus invoke a perceptual shift with these stimuli.

The context bisyllables were recombined with the different steps of the manipulated onset vowels to create the experimental items. This resulted in the following items: for the experimental High F_1 and Low F_1 conditions, instances of [ɪpapu] (deviant), [ˈɛpapu] (standard) and [ɛpapu] (deviant); for the control condition, with the neutral context, instances of [ɪpapu] (deviant), [ɔpapu] (standard) and [ɛpapu] (deviant).

Procedure

Deviant trials were presented randomly mixed between the standards, and presented in the same block (i.e., this was a multi-deviant design). The ratio of presentation was standard 80% and deviants 20% (each deviant 10%). When a deviant occurred it was always followed by at least two standards. The three conditions (high F_1, low F_1 and neutral control conditions) were presented in separate blocks, each one once in the first half of the experiment and once in the second half of the experiment (i.e., each condition was presented twice). Every block consisted of 400 trials (320 standards, 80 deviants). This resulted in a total of 2400 trials. Presentation order of the different blocks was balanced over participants using a Latin-square design.

Participants were instructed to press a button with their right hand whenever they heard a deviant trial. They were instructed that there could be different deviants, and that the difference would always be on the first vowel of the non-words. Participants were also told that a deviant trial would not necessarily sound like a shift in vowel category, but could also be the same vowel that was just pronounced somewhat differently.

Participants were tested individually in a single session in a soundproof, electrically shielded room. They were seated in a comfortable chair at a distance of approximately 60 cm from a computer screen and instructed to relax and avoiding excessive blinking and movements. The instructions were presented on the screen in written form. After each block, participants were allowed to take a break for as long as they wanted. The session included half an hour of electrode application and instruction and one hour stimulus presentation during which the EEG data was recorded.

Button-press responses were measured and analyzed to investigate whether the effect of context resulted in behaviorally measurable differences in detection ability between the different vowels.

EEG recording and analysis

The EEG was recorded from 36 active Ag/AgCl electrodes, of which 32 were mounted in a cap (actiCap), referenced to a nose electrode. Recording and analyses were carried out with Brain Vision Analyzer (version 1.05.0005). Two separate electrodes were placed at the left and right mastoids. Blinks were monitored through an electrode on the infraorbital ridge below the left eye. Horizontal eye movements were monitored through two electrodes in the cap (LEOG and REOG), placed approximately at each outer canthus. The ground electrode was placed on the forehead. Electrode impedance was kept below 20 kΩ, which is a sufficiently low impedance when using active electrodes. EEG and EOG recordings were amplified through BrainAmp DC amplifiers using a bandpass filter of 0.016 - 100 Hz, digitized on-line with a sampling frequency of 500 Hz, and stored for off-line analysis.

Bipolar vertical EOG was computed as the difference between the electrode on the infraorbital ridge of the left eye and the electrode situated right above the left eye. Bipolar horizontal EOG was computed as the difference between the LEOG and REOG electrodes. Data was corrected for the electrooculogram using the Gratton and Coles method in Brain Vision Analyser. EEG was band filtered between 1 and 30Hz

(24 dB/oct). The signals were segmented into epochs of 900 ms (with a 100ms prestimulus baseline, time-locked to the vowel onset). Epochs with an amplitude outside the range of -70 to 70 microvolt were automatically excluded (on average 93.8% of the epochs were kept). Average ERPs were then computed across trials per participant for each type of initial vowel in each type of context.

Wave amplitudes were measured and analyzed for four time windows comprising the time windows P1: 30 - 80 ms (Tavabi, Elling, Dobel, Pantev, & Zwitserlood, 2009), N1: 80 - 160 ms (Cacace & McFarland, 2003), N2: 200 - 300 ms (Celsis et al., 1999; Saarinen, Paavilainen, Schoger, Tervaniemi, & Näätänen, 1992) and P3: 300 - 600 ms (Snyder, Hillyard, & Galambos, 1980). Analyses were conducted on a subset of 15 electrodes (F3, F4, F7, F8, C3, C4, T7, T8, P3, P4, P7, P8, Fz, Cz, Pz). Analyses were conducted separately for the midline (Fz, Cz, Pz) and the lateral (F3, F4, F7, F8, C3, C4, T7, T8, P3, P4, P7, P8) electrodes (see below for how these electrodes were grouped for the analyses).

A first set of analyses was conducted to determine whether an effect of deviant was visible in the ERP. These analyses compared the cortical signatures of the control condition with the neutral context for the standard ([ɔpapu]) versus the average of the two deviants ([ɪpapu] and [ɛpapu]).

A second set of analyses was conducted to compare the average of the two standard stimuli ([ˈɛpapu] in both the High F_1 and the Low F_1 condition) versus the averaged ERPs for the deviants ([ɪpapu] and [ɛpapu] in both the High F_1 and the Low F_1 condition).

The third set of analyses comprised comparisons between the two deviant vowels in the two critical conditions ([ɪpapu] and [ɛpapu] in the high F_1 context condition and [ɪpapu] and [ɛpapu] in the low F_1 context condition). An effect of context on the detectability (reflected in amplitude of the cortical signatures) of the deviant vowels should result in an interaction between the factors Vowel ([ɪ] vs. [ɛ]) and Context (high F_1 vs. low F_1).

Analyses on the electrophysiological data were run using SPSS software. Trials were included in the analysis irrespective of whether they were detected as a deviant or not. We did not differentiate between these so as to maximize statistical power and to keep the number of contributing trials similar across conditions (note in particular that in the critical comparisons we predicted differences across F_1 conditions in the number of deviants that would be detected). The focus of this

research was on the time-course of normalization, and thus not on potential differences between cortical responses to detected versus non-detected deviants, although this question could be the focus of future research. Following the convention for the data analysis of electrophysiological recordings, the Greenhouse-Geisser correction was applied. The analyses investigated the size and location of the effects of deviants in the control and experimental conditions. The lateral analysis (all but the three midline electrodes) included the factors AntPost (with the three levels Anterior, Medial, Posterior), Hemisphere (with the two levels Left and Right), MedLat (with the two levels Medial and Lateral) and Deviant (with the two levels Deviant and Standard). The analysis investigating the size and location of the effect of context on the perception of the vowels (the normalization effect) did not include the factor Deviant but instead included the factors Vowel (with two levels [ɪ] and [ɛ]) and Context (with two levels high F_1 and low F_1). Appendices report the outcomes of the analyses. Effects are only reported if they include the factor Deviant or the interaction between Context and Vowel (for the final, critical, analysis). If interactions were found they were broken up in their constituent (sets of) electrodes. Only highest level interactions were broken down in this way.

Results

Behavioral data

In the control condition participants detected 98.4% of the deviants. In the experimental conditions participants detected on average 52.5% of the deviants. The behavioral data were analyzed using linear mixed-effects models in R (version 2.6.2, R development core team, 2008, with the lmer function from the lme4 package of Bates & Sarkar, 2007). Detection responses were modeled using the logit-linking function (Dixon, 2008). Hits were coded as 1 and misses as 0. Different models were tested in a deductive way, starting from a complete model including the factors Context (levels -1 (low F_1 context) and 1 (high F_1 context)), Vowel (levels -1 (initial [ɪ]) and 1 (initial [ɛ])) and the interaction between Context and Vowel. All factors were centered around zero to make the interpretation of effects more straightforward (Barr, 2008). Non-significant predictors were taken out of the analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. Modelling settled on a main effect for the factor Vowel ($b = 0.080$, $p = 0.002$) which indicates that [ɛpapu] deviants were more easily detected than [ɪpapu] deviants (54.1% vs. 50.9% respectively). An interaction

was found between Vowel and Context (b = -0.586, p< 0.001) which reflects the fact that in the High F_1 condition [ɛpapu] deviants were more easily detected than [ɪpapu] deviants ([ɛpapu] = 65.7%, [ɪpapu] = 37.9%) whereas this effect was in the opposite direction for the Low F_1 condition ([ɛpapu] = 42.6%, [ɪpapu] = 63.9%). The critical normalization effect was thus observed in the behavioral responses.

EEG data: Control condition

Figure 3A displays the grand averages of the standards (dashed) and the deviants (thick solid line) for all 15 electrodes that were included in the analysis, along with the difference waves (thin solid line). Figure 3B displays the average scalp distribution of the difference wave for the 4 different time windows that are analyzed here (a larger set of 28 electrodes were included for the creation of these maps). In general the data show a clear effect of deviant detection, expressed over a large part of the analysis window. Appendix A displays the effects that include the factor Deviant, along with the broken-down constituents for the highest order interactions that include this factor. The four specific time windows will be discussed separately, first describing the analysis of the lateral electrodes and then the midline analysis.

P1 time window (30 - 80 ms).

In the lateral analysis (using all but the three midline electrodes) it was found that there was a negativity for deviants relative to the standard that was significantly expressed over all four posterior electrodes (P7, P3, P4, P8), and only two non-posterior electrodes (C3 and F4). The effect sizes (partial $\eta 2$) were largest on electrodes close to the midline and on posterior electrodes. A similar pattern was found for the midline analysis (Pz, Cz, Fz), which revealed a main effect for deviant vowels along with a two-way interaction with the anterior to posterior dimension. When broken down the latter reflected a significant negativity on Pz, a small effect on Cz, and no significant effect on Fz. This is in line with the posterior distribution of the deviant effects observable from Figure 3b.

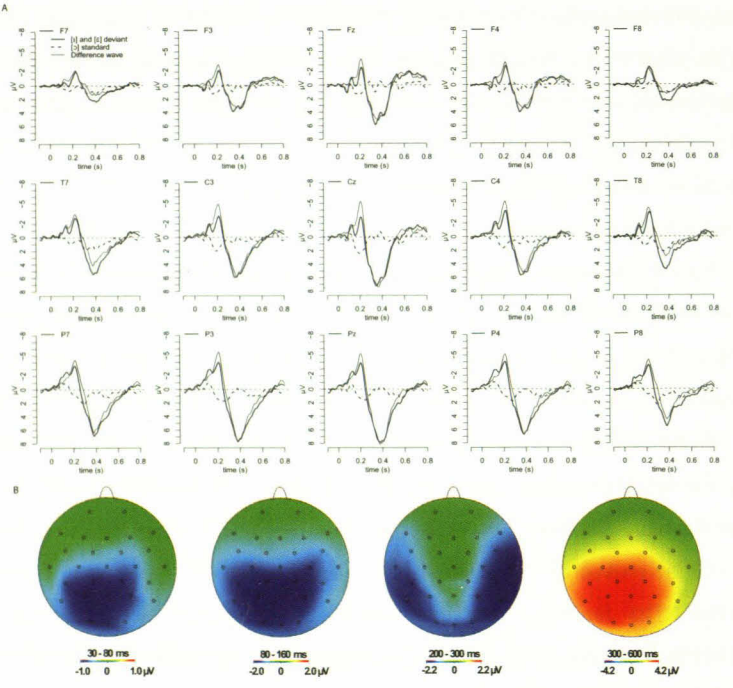

Figure 3. A) Grand averages of the ERPs measured in the control condition for the 15 analyzed electrodes. The dashed line represents the standard stimulus, the thick solid line represents the average over the two deviant stimuli, the thin solid line represents the difference wave between the two (deviant - standard). B) Topographic distributions of the difference wave (average deviant - standard) in the control condition for the 4 selected time-domains P1 (30-80ms); N1 (80-160ms); N2 (200-300ms); P3 (300-600ms). 28 electrodes are displayed. Note that the scales for the separate time windows differ.

N1 time window (80 - 160 ms).

In the lateral analysis, deviant vowels gave rise to negativity effects that were largest towards the midline for posterior and central electrodes, but towards the lateral electrodes for the frontal electrodes. Small hemispheric asymmetries showed that for the right hemisphere negativities were stronger towards midline electrodes while the left hemisphere showed more reliable negativities on the lateral electrodes. The midline analysis revealed a main effect for deviant and an interaction with the anterior to posterior dimension. There were more reliable negativities for posterior electrodes.

N2 time window (200 - 300 ms).

The lateral analysis showed that deviants elicited negativities that were more reliable over lateral electrode sites and towards posterior electrodes. The midline analysis revealed an interaction between the anterior to posterior dimension with the factor Deviant, but when broken down none of the electrodes showed significant effects of Deviant.

P3 time window (300 - 600 ms).

The lateral analysis showed a small hemispheric difference in positivity which showed that for the right hemisphere effects were stronger towards midline electrodes while for the left hemisphere effect were stronger on lateral electrodes. Furthermore, for anterior electrodes effects were stronger over the right hemisphere while for medial electrodes effects were stronger for the left hemisphere. For posterior electrodes effects were of a similar size for the left and right hemisphere. The midline analysis revealed a main effect of Deviant and more reliable positivities towards posterior electrodes.

Summary.

Effects of presenting a deviant non-word were found as a negativity in the earliest time window that was analyzed (30 - 80ms). The effects in this time window were in a negative direction though, indicating that the detection of deviants did not result in a stronger P1. Effects of deviant detection also resulted in strong negativities in the N1 and N2 time windows. Positive enhancement was found in the P3 time domain. In general, these effects were more reliable over posterior sites. These findings show that our design is capable of producing cortical effects during all time windows reflecting the detection of the deviants [ɛpapu] and [ɪpapu].

EEG data: Experimental Standard versus Deviant Analyses

The second analysis compared the average experimental deviant (average over [ɛpapu] and [ɪpapu] in both the high and the low F_1 context conditions) to the average experimental standard ([ʲapu] in both the high and the low F_1 context condition), following the same protocol as the analysis for the above control condition. Figure 4A displays the grand averages of the standards (dashed) and the deviants (solid) for all analyzed electrodes, along with the difference wave (thin solid line). Figure 4B displays the scalp distribution of the difference wave over the 4 different time windows. Appendix B displays the effects that include the factor Deviant, along with the broken-down constituents for the highest order interactions that include this factor.

From Figure 4A, a first observation can be made that the deviant effects are in general much smaller than the effects observed in the control comparison. This was expected as the acoustic difference between the ambiguous sound [ˈɛ] and the deviants [ɛ] and [ɪ] that were used here was much smaller than the acoustic difference between the standards and deviants used in the control condition. Analyses for the separate time windows investigated at what points in time the effect of deviant detection was observed.

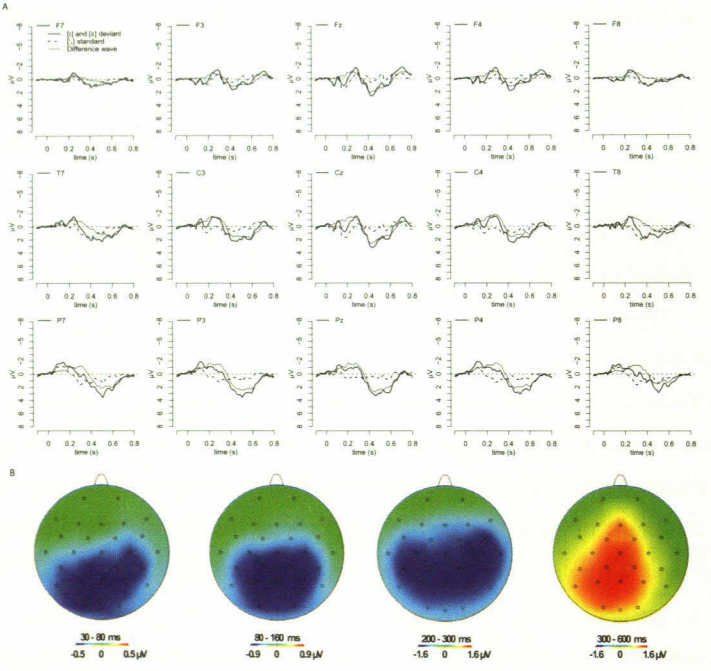

Figure 4. A) Grand averages of the ERPs measured in the experimental condition for the 15 analyzed electrodes. The dashed line represents the standard stimulus (averaged over the two context conditions), the thick solid line represents the average over the two deviant stimuli (averaged over the two context conditions), the thin solid line represents the difference wave between the two (deviant - standard). B) Topographic distributions of the difference wave (average deviant - standard) in the control condition for the 4 selected time-domains P1 (30-80ms); N1 (80-160ms); N2 (200-300ms); P3 (300-600ms). 28 electrodes are displayed. Note that the scales for the separate time windows differ.

P1 time window (30 - 80 ms).

In the analysis of the lateral electrodes (12 electrodes: all but the three midline electrodes) stronger effects of deviants were found towards central and posterior electrodes. The analysis of the midline electrodes (Pz, Cz, Fz) also revealed a main effect and an interaction with the anterior to posterior dimension. When this interaction was broken down effects were stronger towards posterior electrodes, with the strongest effect for Pz.

N1 time window (80 - 160 ms).

The analysis of the lateral electrodes revealed that all but one set of electrodes (F7, F8) revealed significant effects of deviants. The strongest effects were located on posterior electrodes towards the midline (P3, P4). The midline analysis revealed a similar pattern as a main effect of Deviant and an interaction of Deviant with the anterior to posterior dimension was found. When broken down this also showed that the largest negativities were located towards the posterior electrodes (Cz, Pz).

N2 time window (200 - 300 ms).

The analysis of the lateral electrodes revealed that deviants elicited negativities that were spread over all of the analyzed sets of electrodes. The strongest effects were observed, however, over central and posterior electrodes towards the midline (C3, C4, P3, P4). The midline analysis revealed a similar pattern. A main effect of Deviant and an interaction of Deviant with the anterior-posterior dimension was again found. When this was broken down the strongest effects resided on the medial and posterior electrodes (Cz and Pz).

P3 time window (300 - 600 ms).

The analysis of the lateral electrodes revealed effects of Deviant that, when broken down, were stronger over left than over right electrodes, and generally stronger towards the midline. The midline analysis revealed a main effect of Deviant and a small interaction of Deviant with the anterior to posterior dimension. The effect of deviant was strong over the whole midline but slightly stronger towards anterior electrodes.

Summary.

These analyses showed that an effect of the detection of a deviant can also be observed with the relatively small vowel change of the ambiguous sound towards [ɪ] or [ɛ] which were only detected as deviant on 52.5% of the trials. Effects of deviant

detection were again observed mainly over posterior electrodes and during all four tested time windows. We can now test which of these effects were influenced by context.

EEG data: Context effects

Figure 5A displays the difference waves for the vowels presented in the different context conditions (dashed = ([ɪ] low F_1) - ([ɪ] high F_1); solid = ([ɛ] low F_1) - ([ɛ] high F_1)), along with the difference wave between the two (thin solid line). The latter reflects the numerical interaction effect. Figure 5B displays an enlarged version of the panel for electrode P7 (posterior left lateral electrode). Figure 5C displays the scalp distribution of the interaction effect over the 4 different time windows. The behavioral data showed that listeners found it hard to detect a shift from the ambiguous sound [ˈɛ] to {ɪ} in the high F_1 context condition, and from [ˈɛ] to {ɛ] in the low F_1 context condition. If the compensation effect were due to a late and high-level process, the interaction should only be reflected in a late cortical signature such as the P3. If it were due to an early process then it should be observed in an earlier time window such as P1 or N1.

The effect of context should be expressed as follows, taking the P3 response as an example: listeners found it hard to detect a shift from the ambiguous sound [ˈɛ] to [ɪ] in the high F_1 context condition. This should thus result in a small cortical deviant effect in the P3 for that condition. In the low F_1 context condition the detection of a shift from the ambiguous sound [ˈɛ] to [ɪ] should be relatively easier, however, leading to a larger effect on the P3. The difference between these (([ɪ] low F_1) - ([ɪ] high F_1)) was thus expected to result in a positive P3 effect. For the [ɛ] deviants this pattern was expected to be in the opposite direction. The difference line for the [ɛ] deviant (([ɛ] low F_1) - ([ɛ] high F_1)) should therefore result in a negative P3 effect. It was thus expected that the difference line for [ɪ] and the difference line for [ɛ] would show mirror-image patterns around zero. Critically, this relation should hold from the point where the context effect exerts its influence (i.e., not only in the P3 window). This approach therefore shows at what point in time the interaction effect is visible. Figure 5A (and Figure 5B for electrode P7) displays the ERPs and shows the mirror-image pattern in the P3 window, especially on central and posterior electrodes, but also in earlier time-windows. Appendix C reports the significant effects that include the interaction between Context and Vowel (instead of the factor "Deviant" as in the two previous analyses), along with the broken-down highest interactions.

P1 time window (30 - 80 ms).

In the lateral analysis (12 electrodes: all but the three midline electrodes), a small four-way interaction that included the Context by Vowel interaction was found. When broken down, however, none of the separate electrode pairs revealed a significant effect of Context by Vowel. No effect was found in the midline analysis (Pz, Cz, Fz).

N1 time window (80 - 160 ms).

In the lateral analysis two three-way and one four-way interactions that included the critical Context by Vowel interaction were found. When broken down these analyses revealed effects of Context by Vowel only on the left lateral electrodes (F7, T7, P7), and a trend towards an effect on the posterior electrodes (P7, P3, P4, P8; $p = 0.063$, partial $\eta 2 = 0.134$). No effect was found for the midline electrodes.

N2 time window (200 - 300 ms).

The analysis of the lateral electrodes revealed a two-way interaction between Context and Vowel and a three-way interaction of Context by Vowel with the anterior to posterior distribution. When broken down, the strongest effects of Context by Vowel were found over the posterior electrodes (P7, P3, P4, P8). The midline analysis revealed a small Context by Vowel interaction.

P3 time window (300 - 600 ms).

The lateral analysis revealed a three-way and a four-way interaction. When broken down none of the pairs of electrodes reached significance. The midline analysis revealed an interaction of Context by Vowel.

Summary.

These critical analyses revealed cortical signatures of the interaction between the F_1 properties of the /papu/ context and the identity of the target vowel. The earliest reliable effect was observed in the N1 time window (there was a small effect in the P1 time-window, but when broken down none of the electrode sites revealed significant effects). The N1 effect had a left lateralized distribution. The analyses of the N2 and P3 time windows also revealed reflections of the interaction between the F_1 properties of the /papu/ context and vowel identity.

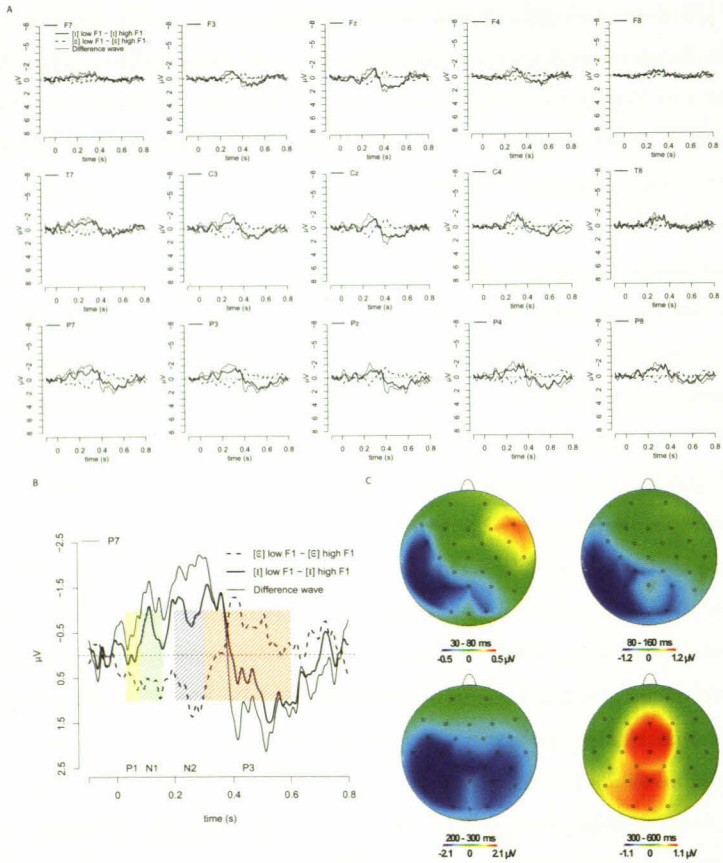

Figure 5. A) Difference waves of the ERP for the 15 analyzed electrodes in the experimental condition, investigating the critical vowel by context interaction. The dashed line represents the difference wave for [ɛpapu] (Low F_1 - High F_1), the thick solid line represents the difference wave for [ɪpapu] (Low F_1 - High F_1), and the thin solid line represents the difference between the two (i.e., the difference between the context effects for each deviant), and thus reflects the interaction effect. B) An enlarged version of electrode P7. The lines represent the same data as in A. Additionally, the four analyzed time windows are indicated along with color coding: P1, yellow (30-80ms); N1, green (80-160ms); N2, blue (200-300ms); P3, red (300-600ms). C) Topographic distributions of the difference wave representing the numerical interaction effect (("Low F_1 [ɛpapu]" - "High F_1 [ɛpapu]") - ("Low F_1 [ɪpapu]" - "High F_1 [ɪpapu])) in the experimental condition for the 4 selected time-domains. Note that the scales for the separate time windows differ.

General discussion

The current paper investigated the level of processing at which compensation for vocal-tract characteristics in speech perception has its influence. In a two-deviant active oddball design, listeners were asked to detect the vowel deviants [ɛ] and [ɪ], embedded in a stream of standards that consisted of an ambiguous sound [ˈɛ] (a sound that was acoustically halfway between [ɛ] and [ɪ]). These standard and deviant vowels were prepended to a non-word context consisting of two syllables (/papu/) that were spliced onto the vowels. Critically, in separate blocks, the /papu/ part was manipulated to mimic either a speaker with a generally high F_1 or a speaker with a generally low F_1 (the standard and deviant vowels were identical across these conditions). These conditions tested compensation for vocal-tract characteristics, as the conditions that mimicked different vocal tracts were expected to influence the detectability of the deviant vowels (represented by both the behavioral and cortical responses) in different ways. In an additional control block the deviant vowels were the same versions of [ɛ] and [ɪ]. However, the standard vowel was [ɔ]. Moreover, in this control condition the /papu/ part had a neutral F_1 range. The cortical responses to these stimuli were recorded and analyzed over 4 different time windows. These were the P1 (30 - 80ms), the N1 (80 - 160 ms), N2 (200 - 300) and the P3 (300 - 600 ms) time windows.

The critical analysis investigated compensation for the F_1 properties of the /papu/ context. The behavioral detection results showed that listeners found it harder to detect a shift from [ˈɛ] to [ɪ] in the high F_1 context condition (compared to detecting a shift from [ˌɛ] to [ɛ]), while the opposite was true in the low F_1 context condition. As has been shown before, this means that the same acoustic token of a vowel, presented in two different contexts, can be perceived differently (Ladefoged & Broadbent, 1957).

The cortical responses to deviants that were recorded during this task showed that the effect of compensation for F_1 properties in the context was not yet observed in the P1 time-domain. A marginally significant 4-way interaction hinted at an effect, but when this interaction was broken down, data from none of the separate electrode sites revealed significant effects. At this level of processing the comparison of the deviant vowels to the representation of the standards did thus not induce reliable differences in the cortical response. During the N1 time window, however, an effect of compensation for context properties was observed. This suggests that the compensation process influences an early level of processing. The negativity in the

N1 time window was observed during the initial vowel and could thus not have been a direct reflection of the exogenous influence of the subsequent context signal (which followed 250 ms after vowel onset).

The fact that we observed an effect in the N1 time domain suggests that the normalization process has a relatively early cognitive locus. As mentioned in the introduction, Roberts et al. (2004) report that a decision bias induced by the identity of a preceding trial did not lead to changes in the N1 of a subsequent trial. We observed the consequences of compensation in the N1 domain and not only during N2 / P3 time windows. This shows that it is unlikely that compensation effects are the result of decision strategies, biases or the listener's conscious interpretation of speech sounds, occurring during the processing of either the standard or the deviant vowel. Instead, normalization appears to change the perception of speech sounds at a very early level of processing.

The control analyses investigated the cortical responses that were recorded during the control condition (standard [ɔ] and deviants [ɛ] and [ɪ], all with a neutral [papu] context) and the cortical responses to the experimental materials (standard: [$^{1}_{\varepsilon}$], averaged over both the high F_1 and the low F_1 context condition; deviants: average over [ɛ] and [ɪ] over both the high F_1 and the low F_1 context condition). These analyses established that our design was capable of eliciting cortical deviant responses, and in which direction (positive or negative) the effects in the different time windows were expected. In both analyses, the deviants resulted in reliable differences in the cortical signature during the P1 time window, observed as a less strong positivity, especially over central posterior electrodes. During the N1 and the N2 time windows, the deviants resulted in stronger negative deflections. During the P1 and N1 time windows the cortical responses to the detection of an oddball were mainly expressed over posterior electrodes. During the N2 time window the distribution seemed to be spreading also towards lateral and frontal areas. This broad negativity had a topographical distribution that was unlike the more frontocentral distribution that is found in classical MMN experiments (Näätänen & Winkler, 1999). This suggests that the negativity should not be readily interpreted as an MMN. This conclusion is supported by the observation that there was no positive deflection at the mastoids during this time window (or the P1 and N1 time windows).

The positivities to deviants in the P3 time domain were similar to other reports of mismatch detection in a response active oddball experimental design. The mainly posterior distribution of the positivity in these conditions suggests that this component should be interpreted as the P3b that is observed after deviants that are infrequent but not unexpected (Friedman, et al., 2001). From a comparison of the deviant waveforms between the midline electrodes (mainly Cz and Pz) in the control condition (Figure 3A) and the experimental standard versus deviant effects (Figure 4A) it can be observed that the P3 signal is longer in the experimental condition than in the control condition. Lengthening of P3 as a function of increasing task difficulty or complexity of the stimulus evaluation is a known characteristic of the P3 wave (McCarthy & Donchin, 1981). Although the deviants [ɛ] and [ɪ] were identical in these conditions the difference between the standard and the deviants was much larger in the control condition (where the standard was the vowel [ɔ]) than in the experimental condition where the standard was the ambiguous sound [$^{1}_{\varepsilon}$] which was acoustically halfway between the deviants [ɛ] and [ɪ]. The two control analyses showed that effects could be observed during all four time windows and indicated whether positivities or negativities should be expected in the different time windows. These analyses validated our analysis of the critical interaction.

The fact that the analyses of the control conditions revealed reliable effects in all time windows allows us to address two questions about the vowel normalization mechanism in the experimental condition: at what point in time does the effects of the compensation mechanism reveal its influence, and at what point in time is that influence finished? With respect to the first question, unlike the analysis of the average standards versus the average deviants, no reliable effect was observed during P1 in the critical experimental analysis. It is important to note here that the critical analysis was based on the analysis of the same (number of) deviant trials as the deviant trials that were used for the analysis of the average standards versus the average deviants. The lack of a reliable effect in the P1 domain can thus be interpreted as evidence that compensation effects affect processes that take place during the N1 time window, and no earlier. However, Figures 5B and 5C show that during the P1 time window there was already a trend towards the effect that became significant during the N1 time window (a left lateralized negativity). There are two possible interpretations of this observation. First, this trend could reflect of changes in perception that have their origin earlier in the processing stream, but which are not

significant due to a lack of experimental power. This would mean that the interaction effect visible during the N1 time window is a mere consequence of the fact that compensation has influenced levels of representation even earlier in the processing stream (possibly in the periphery of the auditory system). It has been shown that brain-stem level processes can influence hair-cell activity in the cochlea in a suppressive manner, and this mechanism has been argued to improve perception in noisy environments (and with extreme intensities as a protective mechanism against acoustic trauma, Kirk & Smith, 2003; May, Budelis, & Niparko, 2004). It is possible that these projections can operate in a sufficiently sophisticated way to induce context effects (Stilp, et al., 2010). Such a mechanism would be a very peripheral implementation of normalization effects.

A second possible interpretation of the initial trend of an effect on P1 is that it is a predecessor of the process that is strong in N1. The processes that become dominant in the N1 time window might already have been partly active during the P1 time window. The first waves of activity after auditory stimulation reach the primary auditory cortex after 10 to 15 ms (Liegeois-Chauvel, Musolino, & Chauvel, 1991). This makes it possible that a change-sensitive process that produces its peak activity during the N1 time domain is already becoming active in earlier time windows.

Some aspects of our data make the second interpretation more likely. It is possible that deviant detection in the control and experimental standard versus deviant conditions led to early deviant effects in subcortical areas. The fact that the earliest time window that was analyzed (P1) showed robust deviant detection effects in the control condition supports this interpretation. This would mean, however, that if compensation processes also influenced early subcortical representations, then, at the level of the cortex, compensation should have been resolved and should only lead to increased detectability. In the critical analysis, the context by vowel interaction in the N1 time-domain showed a distinctively left-sided distribution, mainly on the lateral electrodes. The slowly rising negativity that was visible during the detection of deviants in both the control condition and the overall comparison of experimental deviants versus standards had a distinctly central posterior distribution. The interaction effect visible in the critical context analysis in the P1 and N1 time domains therefore seems to have a different scalp distribution (one that is more strongly left lateralized) than the effects in the same window for the control and the experimental standard-deviant effects (compare the distributions in the N1 time window of Figures

3B, 4B and 5C). These observations suggest that the interaction wave in the critical comparison is not just a reflection of increased detectability as a result of compensation for context instantiated in the peripheral system. The current data therefore suggest that the compensation mechanism under investigation here starts to influence processing just before or during the N1 time window, and certainly no later than that. Future research could focus on earlier components of the ERP signal to further inform the discussion of what the earliest point in time is where compensation mechanisms influences perception.

With respect to the second question, concerning the point in time when the compensation mechanism has no further influence, the current data also provide information. The effects in the later two time windows (N2 and P3) in the context condition were very similar with respect to their topographies to those observed during the control and experimental conditions. This indicates that once compensation has had its influence, the increase in detectability that is the result of this compensation mechanism just adds to the overall detectability of the deviants. The compensation mechanism does not seem to influence any higher representational levels as the cortical response reflecting compensation no longer displays a topography that is different from regular deviant detection. The current findings thus suggest that at the start of the N2 time window (200 ms after target onset) the compensation mechanism has already exerted its influence, and has little additional influence on processing after that point.

A final question about normalization concerns the way in which it is implemented. It has been proposed that an important part of compensation for vocal-tract characteristics in a preceding carrier sentence is based on compensation for average spectral distributions in context signals (Watkins & Makin, 1994, 1996). Listeners continuously build up a representation of the LTAS of preceding sounds, and interpret subsequent signals relative to that LTAS. Watkins and Makin have argued that the result is that listeners perceive target sounds as if they were inversely filtered for the precursor signal by decreasing the perceptual impact of those frequency regions that were very pronounced in that precursor. Such an operation would thus make listeners sensitive to changing acoustic properties (Kluender, et al., 2003; Kluender & Kiefte, 2006; Stilp, et al., 2010). Compensation for LTAS cannot account for all normalization findings (see, for instance, Johnson, Strand, & D'Imperio, 1999, for phoneme categorization-shifts induced by visual context, and

Sjerps et al., 2011, for the demonstration that normalization did not always occur for non-speech stimuli, even though those stimuli had the similar LTAS relations as matched speech stimuli that did elicit normalization). Given the spectral relations between precursors and targets in the experiment presented here, however, compensation for spectral characteristics is likely to play an important role. The N1 has been shown to be sensitive to differences of F_1 (Diesch, et al., 1996; Roberts, et al., 2004), or F_1 and F_2 (Obleser, Elbert, Lahiri, & Eulitz, 2003; Poeppel et al., 1997; Tiitinen, Makela, Makinen, May, & Alku, 2005), and, in an extensive review, Näätänen and Winkler (1999) suggest that the processes underlying N1 retain information on individual static stimulus features. The proposal by Watkins and Makin (1994, 1996) necessarily assumes some form of storage of the LTAS of preceding input. The suggestion that the processes that underlie the N1 retain information on static stimulus features makes them a good candidate for being responsible for LTAS-based compensation. The current results support this interpretation.

Additionally, it has also been suggested that N1 might be sensitive to the F_1/ F_3 ratio instead of absolute formant values (Monahan & Idsardi, 2010). Similar to the rationale in the current paper, that proposal was made in the light of a mechanism that helps listeners deal with between-speaker variation in formant frequencies, but then in a vowel-intrinsic manner (i.e., relying only on information within the target vowel). In contrast to Monahan and Idsardi (2010) and the other papers mentioned above, the current paper focused on vowel-extrinsic influences on perception, as the set of target vowels were identical across the two context conditions. In speech perception, vowel intrinsic normalization and normalization for vowel-extrinsic acoustic information probably operate in tandem (Johnson, 2005), and extrinsic normalization procedures have the potential to play a large role in overcoming between-speaker variation that is the result of anatomical/physiological differences between speakers (Adank, Smits, et al., 2004). The findings reported here, in combination with previous reports in the literature, suggest that the processes underlying the N1 play an important role in normalizing perceptual input to reduce within-category variability.

The method used in this paper presents a novel approach to research on normalization of vowels for the spectral properties of a context speakers' voice. The results demonstrated that compensation for speaker vocal-tract characteristics can be observed through cortical measures and as such provide information about the level of

representation that this process influences. The consequences of compensation for speaker characteristics were observed as soon as 120 ms after vowel onset. This makes it unlikely that the context-induced shifts in perception as they have been observed by others (Holt, 2005; Ladefoged & Broadbent, 1957; Sjerps, et al., 2011; Watkins, 1991; Watkins & Makin, 1994, 1996) were due to strategic effects or biases based on conscious percepts. Instead, compensation for speaker vocal-tract characteristics is for an important part the result of a mechanism that influences an early stage of processing in speech perception.

APPENDIX A

Control condition Comp.Analys.	Factor				F (*df*)	p	partial η^2
P1							
	Lateral						
	Deviant				10.527 (1, 23)	0.004*	0.314
	AntPost* Hemisphere*MedLat*Deviant				3.798 (2, 46)	0.038*	0.142
	Ant	Left	Lat	Deviant (F7)	1.923 (1, 23)	0.179	0.077
			Med	Deviant (F3)	3.125 (1, 23)	0.090	0.120
		Right	Med	Deviant (F4)	5.966 (1, 23)	0.023*	0.206
			Lat	Deviant (F8)	0.027 (1, 23)	0.871	0.001
	Med	Left	Lat	Deviant (T7)	2.653 (1, 23)	0.117	0.103
			Med	Deviant (C3)	12.059 (1, 23)	0.002*	0.344
		Right	Med	Deviant (C4)	0.538 (1, 23)	0.471	0.023
			Lat	Deviant (T8)	0.500 (1, 23)	0.487	0.021
	Post	Left	Lat	Deviant (P7)	8.950 (1, 23)	0.007*	0.280
			Med	Deviant (P3)	16.522 (1, 23)	<0.001*	0.418
		Right	Med	Deviant (P4)	13.545 (1, 23)	0.001*	0.371
			Lat	Deviant (P8)	7.690 (1, 23)	0.011*	0.251
	Midline						
	Deviant				5.162 (1, 23)	0.033*	0.183
	AntPost*Deviant				19.979 (2, 46)	<0.001*	0.465
	Ant	Deviant (Fz)			0.416 (1, 23)	0.525	0.018
	Med	Deviant (Cz)			5.302 (1, 23)	0.031*	0.187
	Post	Deviant (Pz)			11.740 (1, 23)	0.002*	0.338
N1							
	Lateral						
	Deviant				43.426 (1, 23)	<0.001*	0.654
	MedLat*Deviant				9.835 (1, 23)	0.005*	0.300
	AntPost*Deviant				25.320 (2, 46)	<0.001*	0.524
	Hemisphere*MedLat*Deviant				4.516(1, 23)	0.045*	0.164
	Left	Lat	Deviant (F7, T7, P7)		39.291 (1, 23)	<0.001*	0.631
		Med	Deviant (F3, C3, P3)		38.760 (1, 23)	<0.001*	0.628
	Right	Med	Deviant (F4, C4, P4)		39.057 (1, 23)	<0.001*	0.629
		Lat	Deviant (F8, T8, P8)		30.370 (1, 23)	<0.001*	0.569
	AntPost*MedLat*Deviant				8.727 (2, 46)	0.001*	0.275
	Ant	Lat	Deviant (F7, F8)		35.396 (1, 23)	<0.001*	0.606
		Med	Deviant (F3, F4)		20.425 (1, 23)	<0.001*	0.470
	Med	Lat	Deviant (T7, T8)		38.118 (1, 23)	<0.001*	0.624
		Med	Deviant (C3, C4)		41.862 (1, 23)	<0.001*	0.645
	Post	Lat	Deviant (P7, P8)		31.470 (1, 23)	<0.001*	0.578
		Med	Deviant (P3, P4)		45.479 (1, 23)	<0.001*	0.664
	AntPost*Hemisphere*Deviant				3.634 (2, 46)	0.035*	0.136
	Ant	Left	Deviant (F7, F3)		28.014 (1, 23)	<0.001*	0.549
		Right	Deviant (F4, F8)		24.248 (1, 23)	<0.001*	0.513
	Med	Left	Deviant (T7, C3)		38.230 (1, 23)	<0.001*	0.624
		Right	Deviant (C4, T8)		38.205 (1, 23)	<0.001*	0.624
	Post	Left	Deviant (P7, P3)		39.447 (1, 23)	<0.001*	0.632
		Right	Deviant (P4, P8)		37.116 (1, 23)	<0.001*	0.617
	Midline						
	Deviant				16.755 (1, 23)	<0.001*	0.421
	AntPost*Deviant				18.476 (2, 46)	<0.001*	0.445
	Ant	Deviant (Fz)			6.704 (1, 23)	0.016*	0.226
	Med	Deviant (Cz)			15.461 (1, 23)	0.001*	0.402
	Post	Deviant (Pz)			23.652 (1, 23)	<0.001*	0.507

N2						
	Lateral					
		Deviant			15.866 (1, 23)	0.001* 0.408
		MedLat*Deviant			4.611 (1, 23)	0.043* 0.167
		Med	Deviant	(F3, F4, C3, C4, P3, P4)	6.955 (1, 23)	0.015* 0.232
		Lat	Deviant	(F7, F8, T7, T8, P7, P8)	33.340 (1, 23)	<0.001* 0.592
		AntPost*Deviant			9.197 (2, 46)	0.001* 0.286
		Ant	Deviant	(F7, F3, F4, F8)	10.328 (1, 23)	0.004* 0.310
		Med	Deviant	(T7, C3, C4, T8)	15.109 (1, 23)	0.001* 0.396
		Post	Deviant	(P7, P3, P4, P8)	17.445 (1, 23)	<0.001* 0.431
	Midline					
		AntPost*Deviant			3.828 (2, 46)	0.035* 0.143
		Ant	Deviant	(Fz)	0.149 (1, 23)	0.704 0.006
		Med	Deviant	(Cz)	0.026 (1, 23)	0.873 0.001
		Post	Deviant	(Pz)	1.343 (1, 23)	0.258 0.055
P3						
	Lateral					
		Deviant			81.817 (1, 23)	<0.001* 0.781
		MedLat*Deviant			22.720 (1, 23)	<0.001* 0.497
		Hemisphere*Deviant			23.159 (1, 23)	<0.001* 0.502
		AntPost*Deviant			60.290 (2, 46)	<0.001* 0.724
		Hemisphere*MedLat*Deviant			6.244 (1, 23)	0.020* 0.214
		Left	Lat	Deviant (F7, T7, P7)	89.169 (1, 23)	<0.001* 0.795
			Med	Deviant (F3, C3, P3)	58.140 (1, 23)	<0.001* 0.717
		Right	Med	Deviant (F4, C4, P4)	79.908 (1, 23)	<0.001* 0.776
			Lat	Deviant (F8, T8, P8)	61.902 (1, 23)	<0.001* 0.729
		AntPost*MedLat*Deviant			4.214 (2, 46)	0.026* 0.155
		Ant	Lat	Deviant (F7, F8)	41.473 (1, 23)	<0.001* 0.643
			Med	Deviant (F3, F4)	39.373 (1, 23)	<0.001* 0.631
		Med	Lat	Deviant (T7, T8)	89.685 (1, 23)	<0.001* 0.796
			Med	Deviant (C3, C4)	58.754 (1, 23)	<0.001* 0.719
		Post	Lat	Deviant (P7, P8)	73.945 (1, 23)	<0.001* 0.763
			Med	Deviant (P3, P4)	81.729 (1, 23)	<0.001* 0.780
		AntPost*Hemisphere*Deviant			16.433 (2, 46)	<0.001* 0.417
		Ant	Left	Deviant (F7, F3)	35.124 (1, 23)	<0.001* 0.604
			Right	Deviant (F4, F8)	48.262 (1, 23)	<0.001* 0.677
		Med	Left	Deviant (T7, C3)	70.884 (1, 23)	<0.001* 0.755
			Right	Deviant (C4, T8)	64.078 (1, 23)	<0.001* 0.736
		Post	Left	Deviant (P7, P3)	80.429 (1, 23)	<0.001* 0.778
			Right	Deviant (P4, P8)	80.459 (1, 23)	<0.001* 0.778
	Midline					
		Deviant			47.682 (1, 23)	<0.001* 0.675
		AntPost*Deviant			19.404 (2, 46)	<0.001* 0.458
		Ant	Deviant	(Fz)	18.315 (1, 23)	<0.001* 0.443
		Med	Deviant	(Cz)	36.848 (1, 23)	<0.001* 0.616
		Post	Deviant	(Pz)	66.104 (1, 23)	<0.001* 0.742

APPENDIX B

Deviant - Standard Comparison

Comp.	Analys.	Factor			F (*df*)	p	partial η^2
P1							
	Lateral						
		Deviant			11.954 (1, 23)	0.002*	0.342
		AntPost*Deviant			4.818 (2, 46)	0.035*	0.173
		Ant	Deviant	(F7, F3, F4, F8)	4.601 (1, 23)	0.043*	0.167
		Med	Deviant	(T7, C3, C4, T8)	11.404 (1, 23)	0.003*	0.331
		Post	Deviant	(P7, P3, P4, P8)	9.870 (1, 23)	0.005*	0.300
	Midline						
		Deviant			7.237 (1, 23)	0.013*	0.239
		AntPost*Deviant			5.968 (2, 46)	0.014*	0.206
		Ant	Deviant	(Fz)	0.486 (1, 23)	0.493	0.021
		Med	Deviant	(Cz)	7.291 (1, 23)	0.013*	0.241
		Post	Deviant	(Pz)	12.535 (1, 23)	0.002*	0.353
N1							
	Lateral						
		Deviant			31.778 (1, 23)	<0.001*	0.580
		MedLat*Deviant			27.983 (1, 23)	<0.001*	0.549
		AntPost*Deviant			17.830 (2, 46)	<0.001*	0.437
		AntPost*MedLat*Deviant			10.811 (2, 46)	<0.001*	0.320
		Ant	Lat	Deviant (F7, F8)	0.920 (1, 23)	0.347	0.038
			Med	Deviant (F3, F4)	6.802 (1, 23)	0.016*	0.228
		Med	Lat	Deviant (T7, T8)	5.897 (1, 23)	0.023*	0.204
			Med	Deviant (C3, C4)	49.336 (1, 23)	<0.001*	0.682
		Post	Lat	Deviant (P7, P8)	18.125 (1, 23)	<0.001*	0.441
			Med	Deviant (P3, P4)	66.861 (1, 23)	<0.001*	0.744
	Midline						
		Deviant			40.171 (1, 23)	<0.001*	0.636
		AntPost*Deviant			11.154 (2, 46)	0.002*	0.327
		Ant	Deviant	(Fz)	3.678 (1, 23)	0.068	0.138
		Med	Deviant	(Cz)	39.084 (1, 23)	<0.001*	0.630
		Post	Deviant	(Pz)	58.571 (1, 23)	<0.001*	0.718
N2							
	Lateral						
		Deviant			81.526 (1, 23)	<0.001*	0.780
		MedLat*Deviant			15.270 (1, 23)	0.001*	0.399
		AntPost*Deviant			18.705 (2, 46)	<0.001*	0.449
		AntPost*MedLat*Deviant			3.297 (2, 46)	0.050*	0.125
		Ant	Lat	Deviant (F7, F8)	28.098 (1, 23)	<0.001*	0.550
			Med	Deviant (F3, F4)	43.469 (1, 23)	<0.001*	0.654
		Med	Lat	Deviant (T7, T8)	52.219 (1, 23)	<0.001*	0.694
			Med	Deviant (C3, C4)	95.766 (1, 23)	<0.001*	0.806
		Post	Lat	Deviant (P7, P8)	50.073 (1, 23)	<0.001*	0.685
			Med	Deviant (P3, P4)	75.792 (1, 23)	<0.001*	0.767
		AntPost*Hemisphere*Deviant			3.850 (2, 46)	0.030*	0.143
		Ant	Left	Deviant (F7, F3)	25.562 (1, 23)	<0.001*	0.526
			Right	Deviant (F4, F8)	39.471 (1, 23)	<0.001*	0.632
		Med	Left	Deviant (T7, C3)	70.675 (1, 23)	<0.001*	0.754
			Right	Deviant (C4, T8)	58.769 (1, 23)	<0.001*	0.719
		Post	Left	Deviant (P7, P3)	66.837 (1, 23)	<0.001*	0.744
			Right	Deviant (P4, P8)	53.862 (1, 23)	<0.001*	0.701
	Midline						
		Deviant			73.052 (1, 23)	<0.001*	0.761
		AntPost*Deviant			9.705 (2, 46)	0.002*	0.279
		Ant	Deviant	(Fz)	30.336 (1, 23)	<0.001*	0.569
		Med	Deviant	(Cz)	64.379 (1, 23)	<0.001*	0.737

	Post	Deviant (Pz)		83.282 (1, 23)	<0.001*	0.784
P3						
Lateral						
	Deviant			17.046 (1, 23)	<0.001*	0.426
	MedLat*Deviant			21.343 (1, 23)	<0.001*	0.481
	Hemisphere*Deviant			22.806 (1, 23)	<0.001*	0.498
	AntPost*Deviant			15.343 (2, 46)	<0.001*	0.400
	Hemisphere*MedLat*Deviant			7.264 (1, 23)	0.013*	0.240
	Left	Lat	Deviant (F7, T7, P7)	19.616 (1, 23)	<0.001*	0.460
		Med	Deviant (F3, C3, P3)	25.318 (1, 23)	<0.001*	0.524
	Right	Med	Deviant (F4, C4, P4)	15.586 (1, 23)	0.001*	0.404
		Lat	Deviant (F8, T8, P8)	1.823 (1, 23)	0.190	0.073
	AntPost*Hemisphere*Deviant			4.407 (2, 46)	0.018*	0.161
	Ant	Left	Deviant (F7, F3)	14.532 (1, 23)	0.001*	0.387
		Right	Deviant (F4, F8)	5.488 (1, 23)	0.028*	0.193
	Med	Left	Deviant (T7, C3)	23.610 (1, 23)	<0.001*	0.507
		Right	Deviant (C4, T8)	6.190 (1, 23)	0.021*	0.212
	Post	Left	Deviant (P7, P3)	25.934 (1, 23)	<0.001*	0.530
		Right	Deviant (P4, P8)	11.616 (1, 23)	0.002*	0.336
Midline						
	Deviant			21.120 (1, 23)	<0.001*	0.479
	AntPost*Deviant			3.862 (2, 46)	0.047*	0.144
	Ant	Deviant (Fz)		21.350 (1, 23)	<0.001*	0.481
	Med	Deviant (Cz)		19.720 (1, 23)	<0.001*	0.462
	Post	Deviant (Pz)		18.875 (1, 23)	<0.001*	0.451

APPENDIX C

Context Effects

Comp.	Analys.	Factor			F (*df*)	p	partial η^2
P1							
	Lateral						
		Hemisphere *MedLat*Context*Vowel			5.942 (1, 23)	0.023*	0.205
		Left	Lat	Con*Vow (F7, T7, P7)	1.054 (1, 23)	0.315	0.044
			Med	Con*Vow (F3, C3, P3)	0.446 (1, 23)	0.511	0.019
		Right	Med	Con*Vow (F4, C4, P4)	0.010 (1, 23)	0.921	<0.001
			Lat	Con*Vow (F8, T8, P8)	0.532 (1, 23)	0.473	0.023
N1							
	Lateral						
		Hemisphere*Context*Vowel			9.791 (1, 23)	0.005*	0.299
		Hemisphere*MedLat*Context*Vowel			12.934 (1, 23)	0.002*	0.360
		Left	Lat	Con*Vow (F7, T7, P7)	9.118 (1, 23)	0.006*	0.284
			Med	Con*Vow (F3, C3, P3)	3.033 (1, 23)	0.095	0.117
		Right	Med	Con*Vow (F4, C4, P4)	0.234 (1, 23)	0.633	0.010
			Lat	Con*Vow (F8, T8, P8)	0.011 (1, 23)	0.916	<0.001
		AntPost*Context*Vowel			7.944 (2, 46)	0.005*	0.257
		Ant		Con*Vow (F7, F3, F4, F8)	0.004 (1, 23)	0.952	<0.001
		Med		Con*Vow (T7, C3, C4, T8)	0.445 (1, 23)	0.511	0.019
		Post		Con*Vow (P7, P3, P4, P8)	3.831 (1, 23)	0.063	0.143
N2							
	Lateral						
		Context*Vowel			10.294 (1, 23)	0.004*	0.309
		AntPost*Context*Vowel			11.901 (2, 46)	0.001*	0.341
		Ant		Con*Vow (F7, F3, F4, F8)	4.079 (1, 23)	0.055	0.151
		Med		Con*Vow (T7, C3, C4, T8)	11.447 (1, 23)	0.003*	0.332
		Post		Con*Vow (P7, P3, P4, P8)	13.756 (1, 23)	0.001*	0.371
	Midline						
		Context*Vowel			4.478 (1,23)	0.045*	0.163
P3							
	Lateral						
		MedLat*Context*Vowel			10.366 (1, 23)	0.004*	0.311
		AntPost*MedLat*Context*Vowel			5955 (2, 46)	0.009*	0.206
		Ant	Med	Con*Vow (F3, F4)	3.301 (1, 23)	0.082	0.126
			Lat	Con*Vow (F7, F8)	0.563 (1, 23)	0.460	0.024
		Med	Med	Con*Vow (C3, C4)	2.045 (1, 23)	0.166	0.082
			Lat	Con*Vow (T7, T8)	0.030 (1, 23)	0.864	0.001
		Post	Med	Con*Vow (P3, P4)	2.938 (1, 23)	0.100	0.113
			Lat	Con*Vow (P7, P8)	1.452 (1, 23)	0.240	0.059
	Midline						
		Context*Vowel			7.454 (1, 23)	0.012*	0.245

Hemispheric differences in the effects of context on vowel perception.

Chapter 7

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (submitted). Hemispheric differences in the effects of context on vowel perception.

Abstract

Listeners perceive speech sounds relative to context. Contextual influences might differ over hemispheres if different types of auditory processing are lateralized. Hemispheric differences in contextual influences on vowel perception were investigated by presenting speech targets and both speech and non-speech contexts to listeners' right or left ears (contexts and targets either to the same or to opposite ears). Listeners performed a discrimination task. Vowel perception was influenced by acoustic properties of the context signals. The strength of this influence depended on laterality of target presentation, and on the speech/non-speech status of the context signal. We conclude that contrastive contextual influences on vowel perception are stronger when targets are processed predominately by the right hemisphere. In the left hemisphere, contrastive effects are smaller and largely restricted to speech contexts.

Introduction

It is one of the foundations of cognitive neuroscience that language processing relies more on the left than on the right hemisphere. Strokes to the left perisylvian region lead to stronger language impairments than strokes to the right perisylvian region (Ingram, 2007). The Asymmetric Sampling in Time (AST) hypothesis (Hickok & Poeppel, 2007; Poeppel, 2003) accounts for this asymmetry by arguing that the left hemisphere processes acoustic information in a way that is beneficial for the decoding of the speech signal. It is assumed that the left hemisphere integrates information over shorter time windows (i.e., ~20–50 ms) than the right hemisphere (i.e., ~150–300 ms). This makes the left hemisphere well equipped to deal with the fast spectral changes typical for speech sounds, while the right hemisphere is better equipped for fine-grained spectral analysis necessary for the perception of music and intonation contours.

At first glance, this is a very elegant proposal. Scott and Wise (2004), however, argued that it loses much of its appeal once the properties of the human auditory system have been taken into account. Above 1kHz, the auditory system does not allow for fine-grained spectral resolution. Moreover, Scott and Wise argue that there is no convincing evidence that the left auditory cortex has a preference for fast transitions. They conclude that "It is simply not meaningful to consider 'temporal' and 'spectral' in the auditory system as delineating the ends of a dimension which affords rapid temporal resolution at one end and pitch processing at the other" (p. 38).

There may be a different reason, however, why a short window of integration may be useful for speech processing. Short windows are necessary to account for contrastive context effects, such as those first reported by Ladefoged and Broadbent (Ladefoged & Broadbent, 1957). For instance, when participants categorized targets on a continuum ranging from "itch" to "etch", they categorized more stimuli as "itch" when a context sentence was processed by a filter that suppressed the frequencies that are more dominant in ɪ than in ɛ (Watkins, 1991). Similar effects have been observed with non-speech contexts and over relatively long silent intervals between contexts and targets (Holt, 2005; Sjerps, et al., 2011). The common denominator in all these studies is "contrast": A given stimulus is perceived relative to context, so that a "high" context makes "low" percepts more likely, and vice versa (Kluender, et al., 2003). In the case of vowel perception, for example, more vowels on a 1st formant

(F_1) continuum are identified as the low-F_1 endpoint vowel in a context with a high F_1 than in a context with a low F_1.

Contrast effects can obviously only arise if target and context are perceived as separate entities. If information that is processed in the left hemisphere is integrated over shorter time-windows, such that context and target are processed in separate time-windows, contrastive effects should arise ("high" contexts should make "low" percepts more likely). If the right hemisphere, however, uses larger windows of integration, context and target information are more likely to be integrated because they are more likely to fall in the same analysis window ("high" contexts should make "high" percepts more likely). The need to be able to perceive separate acoustic events as separate, a feature that might be especially useful in speech perception, thus constitutes a new raison d'être for the AST hypothesis. This explanation is independent of the motivation based on the distinction between spectral and temporal properties in auditory processing. If this reasoning is correct, we should find contrastive effects for stimuli that are processed primarily by the left hemisphere, but integrative effects for stimuli that are processed primarily by the right hemisphere.

The outcome of different contrastive and integrative effects over the hemispheres could also shed light on some puzzling contradictory findings. As it turns out, the size and direction of context effects have differed across materials. For instance, Watkins (1991) found no effect of contralaterally presented noise contexts on the perception of speech targets, but speech analogs of these stimuli did elicit contrastive effects (Sjerps, et al., 2011). Moreover, integrative effects have been reported in the spectral domain (Aravamudhan, Lotto, & Hawks, 2008; Mitterer, 2006b) and with respect to durational distinctions (Fowler, 1992; van Dommelen, 1999). These inconsistencies between contrastive and integrative effects could reflect differences in the relative involvement of the two hemispheres with speech and non-speech stimuli. The present study was thus set up to test whether hemispheric differences influence extrinsic normalization of vowels. To test this, we made use of two manipulations. First, we used both speech and non-speech stimuli. Second, we presented these stimuli either to participants' right ears or to their left ears.

Monaural input is more strongly transferred to the hemisphere contralaterally to the ear of presentation, for primary and non-primary auditory cortex (Jancke, Wustenberg, Schulze, & Heinze, 2002; Loveless, Vasama, Makela, & Hari, 1994; Stefanatos, Joe, Aguirre, Detre, & Wetmore, 2008; Suzuki et al., 2002). Activation

levels are two to three times as large in the contralateral as in the ipsilateral hemisphere (Jancke, et al., 2002; Suzuki, et al., 2002), although with speech stimuli the contralateral dominance effect has been reported to be larger for the right than for the left ear (Stefanatos, et al., 2008). We manipulated dominance of hemispheric processing by presenting stimuli monaurally to the left or the right ear.

There is, however, a caveat to consider. Signals that are close together in time influence each other at peripheral stages in auditory pathways when presented to the same ear. These influences are contrastive (Summerfield, et al., 1984). Such influences would obscure our investigation because we are interested in central (cortical) levels of processing. Preceding context was therefore separated from targets by a 500 ms silent interval. Moreover, across conditions, contexts and targets were presented either to the same ears or to opposite ears. These precautions allow us to reduce and control the influence of peripheral adaptation (Summerfield, et al., 1984).

We investigated context effects in a 4I-oddity discrimination design. In this task, listeners are asked to detect whether a deviant (D), presented among a set of standards (S), occurred in either second or third position (e.g. SDSS or SSDS). The use of this task reduces influences from response strategies (such as balancing the number of responses between each of the two labels). This is mainly so because the 4I-oddity task does not require the use of category labels, and as such encourages listeners to focus on auditory aspects of target stimuli (Gerrits & Schouten, 2004).

A continuum of target stimuli was created between the Dutch vowels /ɛ/ and /ɪ/ (which is mainly an F_1 distinction). Vowels were presented in a non-word context (/papu/) that was manipulated to have a high- or a low-F_1 contour. A context effect should result in a difference in discriminability between an ambiguous sound with the [ɛ] and [ɪ] endpoints. To exemplify, consider a categorization experiment: In a low-F_1 context, listeners categorize ambiguous vowels more as /ɛ/ (Watkins, 1991). The perceptual distance between the ambiguous sound [$^{\downarrow}_{\varepsilon}$] and [ɛ] is thus smaller in this condition than the distance between [$^{\uparrow}_{\varepsilon}$] and [ɪ]. This pattern reverses for vowels that are presented in a high-F_1 context. In our 4I-oddity discrimination task, context effects should then lead to reduced discriminability between [$^{\downarrow}_{\varepsilon}$] and [ɛ] in a low-$F_1$ context (and between [$^{\uparrow}_{\varepsilon}$] and [ɪ] in a high-$F_1$ context).

Listeners heard sets of three ambiguous standards ([$^{\downarrow}_{\varepsilon}$]) and one unambiguous deviant (either [ɪ] or [ɛ]). The bisyllable [papu] was manipulated to have a high or a low average F_1 and thereby provided listeners with information about the speaker's

vocal tract properties. The context was spliced onto the target vowels such that listeners heard nonsense words like [ˈɛpapu] (standards) and [ɪpapu] or [ɛpapu] (deviants). In one group of listeners the target vowels and contexts were always presented contralaterally. For another group of listeners the target vowels and contexts were always presented to the same ear. The stimuli were presented in sets of four, with the [papu] part identical in all four non-words in a set. Figure 1 displays an example trial for participants in the group that were presented with targets and contexts contralaterally, for a trial in which the targets were presented to the left ear, and with the deviant vowel ([ɛ]) in second position.

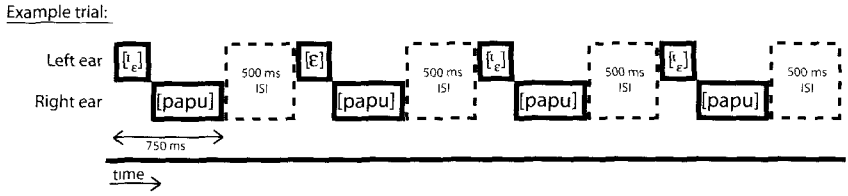

Figure 1: Time-line for an example trial in which context and target are presented to different ears.

In a further condition, the contexts were non-speech stimuli. The [papu] parts now consisted of noise that had the same amplitude envelope as the original [papu] parts. Two of the non-speech versions of the noise precursor were made, one with the same Long-Term Average Spectrum (LTAS) as the low-F_1 [papu] part and one with the same LTAS as the high-F_1 [papu] part. This is important as the LTAS of context signals has been argued to be the main cause of contrast effects (Watkins, 1991).

To summarize, we tested whether contextual influences on vowel perception differ in the two hemispheres. Target vowels were presented in two types of context: a speaker with a high F_1 or a speaker with a low F_1. Effects were tested in a discrimination task. Context effects were expected as a difference in the discriminability of the two deviant vowels across F_1 contexts. Targets were presented to the right or to the left ear (and contexts were, across two groups of listeners, presented to the same and to the opposite ears). Furthermore, context stimuli consisted either of speech (the bisyllable: [papu]) or a non-speech version of this sequence that had the same amplitude envelope and LTAS as the speech version. According to the predictions of the AST hypothesis we should find that contrastive context effects are stronger when stimuli are presented primarily to the left hemisphere (i.e., the right

ear) than when they were presented primarily to the right hemisphere (i.e., the left ear).

Experiment

Method

Participants.

32 native Dutch participants were tested. Participants were invited if they indicated that their right hand was dominant (in response to the question: "Indicate whether you are right or left-handed"). 7 were employees of the Max Planck Institute for Psycholinguistics (MPI) and 25 were participants selected from the MPI participant database (all were uninformed about the purpose of the study). None of the participants reported hearing impairment. All participants can be considered bilingual in at least Dutch and English as the average amount of formal English education for this population is 7–8 years (Broersma & Cutler, 2008).

Stimuli.

For all manipulations Praat was used (Boersma & Weenink, 2005). An instance of [ɛ] was cut out of a recorded version of the non-word [ɛpapu] spoken by a female native speaker of Dutch. Source and filter properties were estimated with Burg's Linear Predictive Coding (LPC) and formant estimation methods in Praat. The F_1 range of the [ɛ] was decreased in three steps over a range of 140 Hz to create a continuum from [ɛ] to [ɪ]. The filter model was then recombined with the estimated source model. The created steps had an F_1 decrease of: -170 Hz ([ɪ]), -100 Hz (ambiguous: [ˡɛ]) and -30 Hz ([ɛ]), relative to the originally recorded instance of [ɛ] (which had an F_1 of 734 Hz). More details on the manipulation procedure can be found in Sjerps et al. (2011). The [papu] context was manipulated with the same approach to create a generally high-F_1 contour (+200 Hz) or a generally low-F_1 contour (-200 Hz). For the non-speech stimuli a noise signal was created that had the same duration and amplitude envelope as an unmanipulated version of the [papu] part. This sound was, in two versions, filtered to match the LTAS of the high-F_1 and of the low-F_1 speech contexts in analogy to Watkins (1991). The vowels were then spliced onto the contexts. Stimuli were combined in quadruplets of three ambiguous standards ([ˡɛpapu]) and one deviant. The deviant could be either [ɛpapu] or [ɪpapu] and deviants occurred in second (SDSS) or third (SSDS) position.

Design.

One group of participants heard only trials in which both context and target were presented to the same ear. A second group always heard targets and precursors in different ears. There were 16 different conditions per participant: ear of target presentation (left, right) by speech status of context (speech, non-speech) by deviant vowel ([ɛ], [ɪ]) by context type (high vs. low F_1). These conditions were presented in separate subparts, which in turn were presented in semi-random order across participants. The order in which the two conditions of a factor were presented was balanced across participants for all factors. Within every subpart three sets of eight trials were presented. Within such a set, the deviant occurred in second or third position four times each, presented in random order. An experiment consisted of 384 trials and lasted 30 minutes. The experiment was divided in four blocks (each containing the subparts for four conditions). Blocks were separated by self-paced pauses.

Analysis and Results

Overall.

The results were analyzed using linear mixed effects models in R (*version 2.10.0*; The R foundation for statistical computing) as provided in the lme4 package (Bates & Sarkar, 2007). For the dichotomous dependent variable of correct responses (i.e., correct = 1 vs. incorrect = 0), a logit linking function was used. Responses were analyzed by fitting models with participants as random factor. All fixed factors were centered around zero. These were Context (with the levels low-F_1 = -1 vs. high-F_1 = 1), indicating the F_1 range in the [papu] part; Deviant (with the levels [ɪ] = -1 vs. [ɛ] = 1), indicating vowel identity; Speech (with the levels Non-speech = -1 vs. Speech = 1), indicating the speech-status of the context; Ear of Target presentation (with the levels Left = -1 vs. Right = 1) indicating the ear to which the target was presented; and Ear of Context presentation (with the levels Left = -1 vs. Right = 1) indicating the ear to which the context was presented. Only those responses that were made after the first vowel of the second target stimulus (the first possible point for mismatch detection) were included (98.8% of the responses were kept).

Figure 2 displays the discrimination scores. Each separate panel displays discrimination scores for the two vowels in the two context conditions. The dotted line represents discrimination scores with a low-F_1 context, the solid line represents the results in the high-F_1 context condition. The values on the left represent the

discrimination scores obtained when the deviant was [ɪ], the values on the right represent those when the deviant was [ɛ]. Context effects are revealed as an interaction between the factors Context and Deviant. The separate panels display the discrimination scores in the different conditions (see the labels in Figure 2).

Consider the top left panel of Figure 2. The pattern reflects that, in the context of a high-F_1 speaker, listeners found it harder to detect a shift from the ambiguous standard [ˈɛ] to [ɪ] (left point of the solid line) than to [ɛ] (right point of the solid line). This is a contrastive influence as the high-F_1 speaker apparently makes the ambiguous stimulus sound more like the low-F_1 vowel (i.e., [ɪ]). This pattern was reversed in the context of a low-F_1 speaker. Analyses were run to test whether this effect was significant and whether it differed over the data for the different panels.

For each analysis, an optimal model was established by a backward-elimination procedure such that non-significant predictors were taken out of the analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was also found. The best model was then established by means of a likelihood-ratio test. The first analysis included all factors along with their interactions.

The optimal statistical model for the overall analysis revealed an effect for the Intercept ($b_{Intercept}$ = 1.153, z = 8.933, p < 0.001) because listeners scored higher than chance. Main effects were found for the following factors: Context ($b_{Context}$ = -0.116, z = -5.344, p < 0.001), indicating that deviant detection was better in the low-F_1 context condition (or its non-speech analog); Deviant ($b_{Deviant}$ = 0.156, z = 7.189, p < 0.001), indicating that deviant detection was better for [ɛ]; Speech (b_{Speech} = -0.144, z = -6.646, p < 0.001), indicating that deviant detection was better when the context consisted of non-speech; and Ear of Context Presentation ($b_{ContextEar}$ = 0.078, z = 3.639, p < 0.001), indicating that deviant detection was better when the context was presented to the right ear. A two-way interaction was found between the factors Context and Deviant ($b_{Context*Deviant}$ = 0.150, z = 6.941, p< 0.001), reflecting the critical context effect which was, on average, contrastive. Finally, two three-way interactions were found between the factors Context, Deviant and Speech ($b_{Context*Deviant*Speech}$ = 0.064, z = 2.965, p = 0.003), and Context, Deviant and Ear of Target Presentation ($b_{Context*Deviant*TargetEar}$ = -0.091, z = -4.222, p < 0.001). Note the direction of these interactions. The positive regression weight for the

Context*Deviant*Speech Interaction indicates that the context effect was more contrastive with speech than with non-speech. The negative regression weight for the Context*Deviant*TargetEar interaction indicates that the context effect was less contrastive when the target was presented to the right ear.

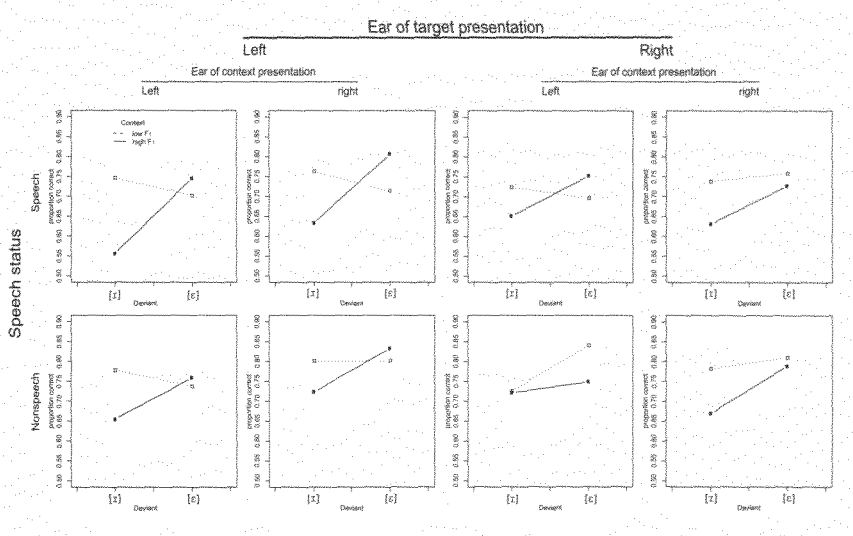

Figure 2. Discrimination data: Mean probability of a correct discrimination response to pairs of stimuli in the 4I-oddity task. Listeners performed a discrimination task in both a Low-F_1 and a High-F_1 speaker condition (defined by the height of the F_1 contour in the [papu] part). Deviant stimuli consisted of either [ɪpapu] ([ɪ]-deviant) or [ɛpapu] ([ɛ]-deviant). The standard stimuli consisted of an ambiguous stimulus [ʔapu]. The eight panels represent the data split over three additional factors: Laterality of presentation of targets; laterality of presentation of contexts; and speech/non-speech status of the context. Left-hand panels display data for which the target was presented to the left ear, right-hand panels display data for which the target was presented to the right ear. Top panels display data for which the contexts consisted of speech signals. Bottom panels display data for which the contexts consisted of noise that had the same amplitude envelope and the same LTAS as the speech contexts (in both the high-F_1 and the low-F_1 conditions).

To further explore this interaction, separate analyses were conducted for the data with right and left target presentation. The optimal models revealed an effect for the Intercept (L: $b_{Intercept}$ = 1.140, z = 9.278, p < 0.001; R: $b_{Intercept}$ = 1.181, z = 8.109, p < 0.001) – reflecting above-chance performance – and the same main effects as the overall analysis (Left: $b_{Context}$ = -0.100, z = -3.282, p = 0.001; $b_{Deviant}$ = 0.144, z =

4.715, $p < 0.001$; $b_{Speech} = -0.149$, $z = -4.886$, $p < 0.001$; Right: $b_{Context} = -0.134$, $z = -4.353$, $p = 0.001$; $b_{Deviant} = 0.171$, $z = 5.575$, $p < 0.001$; $b_{Speech} = -0.148$, $z = -4.829$, $p < 0.001$). The critical two-way interaction between Context and Deviant was found with left target presentation ($b_{Context*Deviant} = 0.244$, $z = 7.949$, $p < 0.001$), and was equivalent for speech and non-speech targets. For target presented to the right, however, the critical interaction was not significant ($b_{Context*Deviant} = 0.059$, $z = 1.914$, $p = 0.056$) and qualified by an additional interaction ($b_{Context*Deviant*Speech} = 0.073$, $z = 2.386$, $p = 0.017$). Additional testing with stimuli presented to the right ear showed that the critical interaction was present with speech contexts ($b_{Context*Deviant} = 0.134$, $z = 3.183$, $p = 0.001$), but not for non-speech contexts ($b_{Context*Deviant} = -0.013$, $z = -0.307$, $p = 0.759$). In the latter case, we observed an integrative effect for targets preceded by an ipsilateral context ($b_{Context*Deviant} = -0.149$, $z = -2.343$, $p = 0.019$). Ipsilateral presentation rules out the possibility of peripheral context effects, which are always contrastive. The latter condition was thus expected to be most sensitive to potential integrative effects.

Discussion

This study was set up to investigate a prediction derived from AST (Poeppel, 2003) with respect to contextual influences on vowel perception. AST proposes that the right hemisphere integrates information over longer time windows than the left hemisphere. This led to the prediction that processes in the right hemisphere would lead to more integrative effects than those in the left hemisphere. Combined with the fact that biological systems are naturally more sensitive to contrast (Kluender & Kiefte, 2006), it was predicted that processing in the left hemisphere would induce stronger contrastive effects on vowel targets than processing in the right hemisphere. This hypothesis offered a possible explanation for earlier demonstrations of variation in the strength and direction of context effects on vowel perception (Aravamudhan, et al., 2008; Mitterer, 2006b; Sjerps, et al., 2011; Wade & Holt, 2005; Watkins, 1991). We indeed observed an influence of laterality on the strength of context effects, but the influence was in the opposite direction from that predicted by AST.

We probed laterality of processing using the fact that transfer of information is stronger over contralateral than ipsilateral connections between the cochlea and the cortex (Jancke, et al., 2002; Loveless, et al., 1994; Suzuki, et al., 2002). In general, contrastive context effects on vowel perception were observed. In the context of a speaker with a high-F_1 contour, listeners found it more difficult to detect a shift from

[ˈɛ] to [ɪ] (a low F_1 vowel) than a shift from [ˈɛ] to [ɛ] (a high F_1 vowel). In the context of a speaker with a low-F_1 contour, this effect was reversed. However, when the target signal was presented to the left ear, so that its most dominant processing would presumably be primarily in the right hemisphere, the contrastive context effect was larger than when the target signal was presented to the right ear. Furthermore, when the context consisted of a speech signal, the contrastive effect was generally larger than when it was replaced by a noise version, as has previously been reported (Watkins, 1991).

The separate analyses of targets presented to the two ears showed that, for targets presented to the left ear, context effects were rather uniformly distributed, such that the factor of speech status did not significantly modulate the strength of context effects. This suggests that the right hemisphere operates in a generally contrastive way, which is a natural tendency of biological systems (Kluender & Kiefte, 2006). In the left hemisphere, however, the tendency for contrastive effects was strongly reduced. In particular, the contrastive effect due to predominately left-hemisphere processing was found for speech stimuli but not for non-speech stimuli. We suggest that, especially for the left hemisphere, context effects could be more dependent on learnt properties of language. Exposure to language, and the learnt covariations between vocal tract properties within a particular speaker, might have influenced the tendency for listeners to compensate for vocal tract properties in a preceding sentence.

This suggestion, however, still does not completely explain the conflicting results reported in the literature that were the starting point of this series of experiments (Aravamudhan, et al., 2008; Sjerps, et al., 2011; Wade & Holt, 2005; Watkins, 1991), such as the occurrence of integrative effects. In the experiment reported here, variance in the strength of context effects was also found, and this was mainly predictable on the basis of the speech status of the contexts and the hemisphere in which the target was most dominantly processed. In one condition we observed a small integrative effect, one that was not very reliable. The same combination of the factors ear of target presentation and speech status of the context led to a contrastive effect when target and context where presented to the same ear (compare the two bottom-left panels of Figure 2). Given the occurrence of integrative effects reported in the literature and here, however, we think it is still tenable that there is an interplay between contrastive and integrative effects. More research

specifically aimed at differences in the strength and laterality (rather than occurrence) of context effects is necessary.

The data presented here show that the two hemispheres contribute differently to context effects. This general observation seems to be consistent with the predictions of AST. The detailed results, however, paradoxically turn out to be opposite from those predicted by AST. Furthermore, an additional observation is also not in line with the AST proposal. If the right hemisphere would be better at resolving spectral differences, target stimuli should be better discriminated when presented to the left ear than when presented to the right ear. No such main effect for ear-of-target presentation was found. In fact, ear of target presentation only modulated the context by target interaction. These findings shows that, for vowel discrimination per se, neither of the hemispheres has an advantage over the other, resonating with Scott and Wise's (2004) criticism of the AST proposal. The current findings support this view, but add an important proviso: The two hemispheres do display different influences of the spectral properties of *contexts*.

The present study is an attempt to reconcile conflicting findings on context effects with AST. Although our data do not support AST, they do show that the different hemispheres contribute to context effects in a different fashion. The results suggest that variability in the strength of context effects is for an important part dependent on the hemisphere in which the target sounds are most dominantly processed. This observation provides a new window into the investigation of central context effects in speech perception. Furthermore, we also provide data that are important in determining differences in hemispheric processing: The two hemispheres may not be differentially sensitive to spectral properties of stimuli per se, but they do show different sensitivities to acoustic properties of context sounds. This is important, for instance because stimuli can elicit context effects on subsequent stimuli. These results thus provide restrictions for the design of future experiments that attempt to investigate differences in hemispheric specialization in speech perception. The present findings already make clear, however, that hemispheric differences do impact on the way vowels are perceived in context.

SUMMARY AND CONCLUSIONS

Chapter 8

Summary of the results

The aim of the series of experiments presented in this thesis was to investigate how listeners manage to deal with variation in speech. There are a number of different processes at work during speech perception that help listeners deal with variation. The current thesis focused on compensation for speakers' vocal tract characteristics. Speakers vary widely in the shapes of their vocal tract and in their speaking style. A consequence of the variance in vocal tracts is that two different speakers producing the same vowel, /ɛ/ for example, can produce very different sounds. In some cases such differences can be so severe that the F_1 and F_2 of the [ɛ] of one speaker might be more similar to the [ɪ] than to the [ɛ] of another speaker. Listeners, however, are able to map these variable signals onto the correct phoneme categories. They do so in part because they compensate for the general vocal tract characteristics of the speakers that they are listening to (Ladefoged & Broadbent, 1957). Ladefoged and Broadbent (1957) showed that characteristics that are revealed in a sentence influence the perception of a following target. This finding formed the basis for the experiments reported here. In the following I will refer to the type of effect that was reported by Ladefoged and Broadbent (i.e., influences of a precursor on the perception of subsequent targets) as "extrinsic normalization" effects.

In Chapter 2 I investigated the speech-specificity of extrinsic normalization effects. I created a set of materials that had undergone a number of manipulations that rendered speech stimuli non-speech-like. The first experiment had two purposes. To replicate the effect originally reported by Ladefoged and Broadbent (1957), and to test whether the effect was specific to speech. Experiment 1 consisted of two parts, a speech and a non-speech part. The speech part was an experiment in Dutch in which participants categorized targets of a [pɪt] to [pɛt] continuum. These targets were preceded by a context sentence ("op dat boek staat niet de naam", lit: on that book is not the name). In two conditions this context sentence was manipulated to have either a high or a low F_1 contour. Participants categorized more sounds as /pɪt/ (which has a

low F_1) when the preceding context had a high F_1 contour than when it had a low F_1 contour (and vice versa for /pɛt/). For the second, non-speech part of Experiment 1, these materials were all spectrally rotated. Spectral rotation changed the spectral pattern of the signal such that information in high frequency ranges traded place with information in low frequency ranges. The resulting signal does not sound like speech. For this experiment listeners were asked to categorize targets from the spectrally rotated versions of the [pɪt] to [pɛt] continuum. When these targets were preceded by spectrally rotated versions of the high- and low- F_1 context sentences, compensation effects were again found. When listeners hear sounds that they do not perceive as speech they still compensate for the spectral characteristics of the preceding context.

In Experiment 2 the precursor materials of Experiment 1a were manipulated more extensively to make them even more unlike speech. This was achieved with the following five manipulations: (1) the materials were spectrally rotated; (2) the materials were given a monotone pitch contour; (3) silent intervals (such as those due to stop closures) were removed; (4) all syllables were temporally reversed; and (5) all syllables were given an overall equal amplitude. When participants categorized the spectrally rotated speech sounds preceded by these strongly manipulated materials, no normalization effects were found. This showed that not all sounds induce normalization effects. This finding contradicts the suggestion that the amount of normalization that is observed is solely dependent on the relation between the Long Term Average Spectra (LTAS) of precursor and target signals (Watkins, 1991; Watkins & Makin, 1994), because the LTAS relationships in this experiment were equivalent to those in Experiment 1.

Further experiments tested for normalization with other combinations of these five manipulations. It was found that the only two manipulated signals that gave rise to normalization effects were in a way opposites: Experiment 1b (spectrally rotated speech) and Experiments 4c and 4d (all manipulations *except* for spectral rotation). These findings show that there is not a single acoustic aspect that triggers the normalization effect. Rather, it seems that general acoustic similarity to speech determines whether normalization effects apply. A final experiment investigated whether it was the perceptual similarity to speech that determined whether normalization occurred or not. In this experiment, participants rated how speech-like the stimuli sounded. There was no clear correspondence between the perceived speechiness and the amount of normalization that stimuli had induced.

CHAPTER 8: SUMMARY AND CONCLUSIONS

In Chapter 3 it was investigated whether attention could have had a modulating influence on normalization. It was possible that the non-speech precursors in Chapter 2 had failed to induce normalization effects because the listeners did not pay attention to them. Listeners might have felt that the extremely non-speech materials were somehow less relevant for the recognition of the subsequent target and could have tried to ignore the non-speech stimuli more than the speech stimuli. To directly investigate the influence of attention, an experiment was run in which participants were encouraged to pay attention to the precursors by means of an additional task. For this task listeners had to refrain from responding to the targets whenever they heard that the precursor contained a dip in amplitude. To test whether the attentional manipulation had an influence on the size of normalization effects, the same materials were used as those from Experiments 1b and 4c in Chapter 2. These were the precursor-target sets that elicited normalization effects although these materials had been manipulated. It was found that the attentional manipulation did not increase the normalization effect. This finding suggests that the strength and direction of the normalization effect are mostly dependent on signal characteristics.

In Chapter 2 it was suggested that linguistic exposure could have had an influence on extrinsic normalization effects. Chapter 4 investigated whether differences in the amount of normalization could be observed in different languages to test this proposal. An experiment was run with very similar stimuli to those of Experiment 1a of Chapter 2 (an experiment with spoken Dutch materials). Stimuli were constructed in English, Dutch and Spanish and categorized by native speakers of English, native speakers of Dutch and native speakers of Spanish. For the native speakers of Spanish, two groups were tested. One group had low proficiency in English; the other group were Spanish-English bilinguals. Participants categorized targets on a continuum from [sufu] to [sofo] (/o/ and /u/ are shared among the stimulus languages). The vowels of this continuum are mainly distinguished by the height of their F_1. These targets were preceded by context sentences that were manipulated to have a high or a low F_1 contour. The across-language design made it possible to investigate whether the size of normalization effects between listeners with different language backgrounds depended on listeners' familiarity with a language, the importance of F_1 as a cue in a particular language, the number of monophthongal vowels in a listeners' mother tongue, or whether any differences in the amount of

normalization obtained with speech sounds depends only on the LTAS relation between precursor and targets.

The results showed that listeners of all language backgrounds compensate for the vocal tract properties of a speaker in a preceding sentence. They did so irrespective of the language in which this sentence was uttered. However, a complex pattern of the amount of compensation across the sets of materials was also found. Spanish listeners compensated most strongly of all listener groups and did so for all materials. Dutch listeners, in contrast, compensated a lot for materials that were presented in Dutch, but less for materials in English and least for materials presented in Spanish. The bilingual listeners compensated a little less than the native speakers of Spanish while the English listeners compensated to a similar extent as the native speakers of Spanish. Furthermore, analyses that were performed separately for each language background group indicated that the English and the bilingual listeners showed differences in the amount of compensation across the materials as well.

Although there is not a single and straightforward explanation for the complete pattern of these results, a number of important conclusions can be drawn. First, it was shown that all listener groups compensated for vocal tract characteristics in a preceding sentence, even when that sentence was spoken in a second or completely unfamiliar language. Thus, compensation for vocal tract characteristics is a mechanism that applies to speech signals in a relatively general way. Second, even within the set of speech sounds, differences in the amount of normalization occurred. These findings thus add to the results of Chapter 2 that variation in the amount of normalization is found not only between speech and non-speech sounds but also among speech sounds from different languages. Listeners from different language backgrounds have learnt to pick up on different perceptual cues in speech signals. The findings of Chapter 4 suggest that such different sensitivities also influence the strength of the impact that a precursor signal has on a subsequent target.

Chapter 5 reported on a combination of categorization and discrimination tasks probing extrinsic vowel normalization. Previous research on extrinsic vowel normalization had used only categorization tasks. However, categorization tasks encourage listeners to focus on categorical rather than on pre-categorical properties of stimuli. A dominant view on extrinsic normalization suggests that an important part of extrinsic vowel normalization arises at a pre-categorical level of processing (Kluender & Kiefte, 2006; Sjerps, Mitterer, & McQueen, 2011 [Chapter 2]; Stilp, Alexander,

Kiefte, & Kluender, 2010; Watkins, 1991; Watkins & Makin, 1994, 1996). To investigate whether normalization does indeed have a strong pre-categorical component, I decided to test whether normalization effects could also be observed in a discrimination task which is argued to reflect auditory processing levels (Gerrits & Schouten, 2004). A set of stimuli was first created that contained vowels on an [ɪ] to [ɛ] continuum. These vowels were spliced onto a [papu] context to form a non-word continuum from [ɪpapu] to [ɛpapu]. Crucially, in two conditions, the [papu] part was manipulated to have either a high or a low F_1 contour. In a first experiment, participants categorized targets from these continua. This showed that listeners indeed perceived an [ɛpapu] to [ɪpapu] continuum and that their responses were influenced by the F_1 contour in the [papu] context. This contextual influence was in the same direction as in the experiments reported in the previous chapters.

In the next experiment these stimuli were used in a 4I-oddity discrimination task: Listeners heard a sequence of 4 stimuli of which either the 2nd or 3rd stimulus was a deviant (AABA vs. ABAA). Participants had to indicate which stimulus was the deviant one. More specifically, participants discriminated standard [ˈɛpapu] stimuli with an initial ambiguous sound [ˈɛ] (perceptually halfway between [ɪ] and [ɛ]) from deviant stimuli [ɪpapu] and [ɛpapu]. As in the previous experiment, the [papu] contexts had either a high F_1 or a low F_1 contour. If a perceptual difference between two stimuli is smaller, then this should lead to worse discrimination scores (fewer correct answers). I predicted that if normalization is due to a process that operates at an auditory processing level, discrimination scores along the [ɪpapu] to [ɛpapu] continuum would depend on the F_1 contour in the speaker context. It was found that discrimination scores did indeed depend on the F_1 context, and that the context influenced perception of the target vowels in a contrastive way. A final experiment confirmed that the 4I-oddity task had indeed encouraged listeners to focus on auditory aspects of the stimuli. These experiments showed that perceptual shifts in vowel perception are instantiated at pre-categorical levels of processing.

Chapter 6 reported on an electroencephalography (EEG) experiment that was set up to gain a better insight into the neural correlates of the normalization process and, more specifically, its temporal development. For this investigation the same stimuli as those used in the experiments in the previous chapter were used. Listeners were asked to listen to a stream of non-words and press a button whenever they heard a deviant. The standard sounds again consisted of the ambiguous non-word [ˈɛpapu]

while the deviant was either [ɛpapu] or [ɪpapu]. Again, the [papu] part had either a high or a low F_1 contour. It was predicted that the different types of context would lead to different detectability of a deviant. In addition, the EEG design allowed us to monitor at what point in time the normalization process has its influence.

Similar to the behavioral results presented in Chapter 5, listeners found it harder to detect a shift from [ˈɛ] to[ɪ] when the F_1 contour of the [papu] part was high while they found it relatively harder to detect a shift from [ˈɛ] to [ɛ] when the F_1 contour in the [papu] part was low. This interaction was also found in the event-related potentials (ERPs) of the associated cortical responses. The ERPs showed that the processes underlying normalization were already having an effect after ~120 ms (in the N1 time window: 80 - 160 ms). Compensation processes therefore seem to operate during a relatively early time window. No reliable effect of normalization was found in an earlier time window though (the P1 time window: 30 - 80 ms). This suggests that normalization does not operate before the N1 time window.

In the final experimental chapter (Chapter 7), I investigated whether normalization effects differ across the two hemispheres. Normalization effects were investigated for speech and non-speech contexts and for targets presented to the right or to the left ear. Sound presented to one ear is most predominantly processed in the contralateral hemisphere (Kimura, 1961; Stefanatos, et al., 2008). This allowed us to manipulate the hemisphere in which the target vowels were most dominantly processed. The stimuli were again very similar to those used in Chapters 5 and 6. Listeners discriminated between [ˈɛ] and [ɪ] or between [ˈɛ] and [ɛ] in a 4I-oddity design. Again, the [papu] context was manipulated to have either a high or a low F_1 contour (or, for the non-speech contexts, noise with the same LTAS as those of the two F_1 speech conditions). The size and direction of compensation was strongly dependent on the mode of presentation. Overall I found that non-speech contexts had a smaller influence on subsequent vowels than speech contexts (similar to the results reported in Chapter 2). However, I also observed an effect of laterality of processing. The more strongly contrastive normalization effects were observed for targets presented to the left ears, which were presumably more strongly processed in the right hemisphere. In one condition, an integrative effect was observed. Although this effect was very small it suggests that variation in the amount of normalization that is observed across different materials might be due to competing forces (integrative

versus contrastive). Further research will have to address this issue more thoroughly though.

Conclusions

When the findings of this thesis are integrated a number of important aspects of extrinsic vowel normalization can be identified. First, it was asked when normalization operates in the hierarchy of speech processing. A second aspect, dealing with possible hemispheric differences, addresses the proposal that the processing of context signals in the different hemispheres has different influences on vowel perception. Third, I have gathered evidence that concerns the prerequisites for extrinsic normalization. This evidence addresses the question whether all sounds undergo normalization in a similar way. These issues will be discussed in turn. In a final section I will discuss how extrinsic normalization combines with other compensation mechanisms to resolve variability in speech.

Cognitive locus

An important contribution of these findings to the literature on extrinsic normalization is in the information they provide about the point in the processing hierarchy at which normalization takes place. To address when normalization starts, however, a distinction needs to be made between the normalization process and a process that is functionally different but that can have a similar influence. In the experiments reported in this thesis I have tried to focus on compensation for context that is caused by "central" instead of "peripheral" influences. An example of a peripheral influence is the fact that a one-second long acoustic segment (the precursor) whose spectrum contains valleys in places where a vowel has peaks and vice versa (i.e., the complement of a vowel) can induce the percept of its complement vowel on a subsequent noise sound that has a uniform spectrum. But this effect disappears when the precursor is shorter than 150 ms, when a silent interval of more than 500 ms separates the precursor and noise-target, or when precursor and the noise-target are presented to different ears (Summerfield, et al., 1984). This shows that there is a class of contextual influences that are short-lived and do not influence the perception of stimuli that were presented to the other ear. These influences were defined as "peripheral influences". In this thesis I focused instead on "central influences". Central influences can be observed across longer intervals and after contralateral presentation (e.g., Chapters 2 through 7; Watkins, 1991). This shows that the effect reported by Watkins (1991), and the effects reported here are functionally

different from those reported by Summerfield, et al. (1984). The dissociation between peripheral and central influences was strengthened by the findings of Chapter 6. No influence of normalization was found during the P1 time window (30-80 ms), while two control experiments showed that our design was in principle capable of measuring differences during that time window. This suggests that normalization effects do not influence the perception of a vowel before ~80 ms after onset.

The results of Chapter 6 did reveal effects in the 80 to 160 ms time window. The most dominant peak during this time window, the N1, has been argued to reflect processes that operate on the boundary between auditory and phonemic processing (Näätänen & Winkler, 1999; Roberts, et al., 2004; Tavabi, et al., 2007). It is therefore likely that the normalization process influences auditory perception at or just before the extraction of phonemic cues. This conclusion is consistent with the findings in Chapter 5. In that chapter normalization effects were observed in a task that encouraged listeners to focus on pre-categorical properties of the stimuli.

Lateralized processes?

In Chapter 7 laterality of processing of context information was manipulated. An influence was observed of ear-of-presentation. The results suggested that the right hemisphere induces contrastive effects on vowel perception, both when the contexts consisted of noise and when they consisted of speech sounds. In the left hemisphere, however, contrastive effects were only found with speech contexts. Based on the results in Chapters 2 and 4, it was proposed that linguistic exposure could have caused differences in the impact of precursors. The finding reported in Chapter 7 suggests that such differences could, for an important part, reside in the left hemisphere. The strength and direction of vowel normalization under binaural presentation conditions is then probably the result of these two influences, and the extent to which a task relies on lateralized processing.

As such, the findings presented in Chapter 7 also speak to the Asymmetric Sampling in Time (AST) hypothesis (Poeppel, 2003). This framework suggests that the left and right hemisphere integrate information over different time windows. If, as proposed by AST, the right hemisphere integrates information over longer time windows, an integrative influence of context would be expected. Due to the longer window of integration, the different speech sounds are perceived as one object. For the left hemisphere, vowel target and context are not in the same window of integration, so instead they are perceived as separate objects. Paradoxically, the

finding of Chapter 7 were almost completely opposite from those predicted by AST. The strongest *contrastive* effects were observed for targets that were more dominantly processed in the right hemisphere.

An additional finding was that in one condition, with targets presented to the right ear and non-speech context to the left, I observed an integrative effect. This suggests that the lack of an effect in the left hemisphere with non-speech sounds could be due to conflicting forces (i.e., contrastive and integrative effects). This hypothesis is interesting as it provides a link to studies that report variable amount of normalization (and sometimes even integrative effects) with non-speech sounds (Aravamudhan, et al., 2008; Fowler, 1992; Mitterer, 2006b; van Dommelen, 1999). However, while this provides an interesting lead into explaining some of the mixed findings on context effects in speech perception, more research is needed. First, it will be necessary to measure more directly to what extent the ear-of-presentation manipulation that I used actually causes context signals to be processed more on the right or more on the left through, for instance, fMRI measures. Second, I observed the integrative effect in only one condition so it is necessary to replicate this particular finding.

Prerequisites for normalization: speech specificity?

In this thesis I explored a number of possible prerequisites for extrinsic normalization. The first possibility was that only speech sounds could elicit normalization effects. This hypothesis was disconfirmed by the finding of strong normalization effects with noise contexts (when the target was presented to the left ear, Chapter 7), and the finding of normalization effects with spectrally rotated signals in Chapter 2.

A second hypothesis would be that there are no prerequisites and that all precursors induce normalization effects. This hypothesis was also disconfirmed, however, because not all non-speech precursors induced normalization effects (Chapter 2). Moreover, the same precursor signal was found to induce different amount of normalization depending on a listener's background language (Chapter 4). These findings disagree with previous suggestions in the literature. For instance, Watkins (1991) has suggested that LTAS relations determine the strength of normalization effects. Watkins argued that the lack of normalization effects with noise that has no spectrotemporal variation is a special case in this respect (i.e., that the only stimuli that do not induce normalization are those without spectrotemporal variation).

I disconfirmed that conclusion because I observed a lack of normalization with signals that did have spectrotemporal variation (Experiment 2 of Chapter 2).

The third possibility was that the size of the effects is dependent on the subjective speech-like qualities and/or the amount of attention that a listener pays to a precursor signal. These interpretations, however, were disconfirmed. In the rating experiment reported in Chapter 2, it was found that signals that induced stronger normalization effects were not necessarily the signals that were rated most speechlike. In a related experiment, Watkins & Makin (1994) came to a similar conclusion. In their experiment participants rated how "natural" a number of voices sounded. As with the "speechiness" ratings in Chapter 2, they also failed to find a relation between subjective qualities and the size of normalization effects. Furthermore, a task that encouraged listeners to pay overt attention to the precursors did not increase the amount of normalization (Chapter 3). These combined findings suggest that the strength of the normalization process is for the most part dependent on bottom-up signal characteristics, and not on attention or subjective qualities.

The fourth possibility that was investigated was that a single acoustic property exists that can solely trigger normalization effects. The strongly manipulated precursor used in Experiment 2 in Chapter 2 failed to induce a normalization effect. Speech stimuli, on the other hand, have reliably elicited compensation effects on multiple occasions in the literature and in this thesis. In Chapter 2, two individual characteristics, pitch movement (Experiment 4a) and the occurrence of low amplitude parts (Experiment 4b) were applied to the non-speech signal that failed to induce an effect in Experiment 2. Neither of these aspects turned out to be the crucial acoustic property. Others have reported compensation effects with temporally reversed speech (Watkins, 1991), suggesting that reversing syllables is not a crucial factor either. It thus seems that none of the properties that were individually manipulated in Chapter 2 triggered normalization. This suggestion is strengthened by the fact that the two signals that were in a way opposites, those in Experiment 1b (only spectral rotation) and those in Experiment 4c (all but spectral rotation) both elicited normalization effects.

I suggest that it is more likely that normalization is triggered by acoustic similarity to speech in a more broadly defined way, such that a number of combined cues can make a signal speech-like and hence induce normalization, but also such that different combinations of cues can do so. This adds to the idea that normalization is a

graded process. In Chapter 4 (normalization with different languages) I found that, even among all-speech stimuli, the strength of the normalization effect can vary. Normalization effects therefore depend in some way on the specific linguistic background of a speaker in combination with the stimulus language a participant is listening to. Although the exact pattern of differences in the size of the normalization effect across stimuli in different languages were unclear, it shows that familiarity to the stimuli influences the strength of normalization. This observation aligns with a suggestion made in the discussion section of Chapter 2. Listeners might learn through exposure to language that it can be beneficial to compensate for spectral characteristics available in context. The amount of compensation with a particular set of materials is then dependent on the exact type of exposure that a listener has had. Speakers of different languages are naturally sensitive to slightly different speech cues. This is a result of the differences in phonemic properties in those languages. The results of Chapter 4 suggest that differences in sensitivity to such cues influence the impact that a precursor has on a subsequent speech sound. A question for further investigation is how the exposure of Dutch listeners differs from that of, for example, the English participants so as to explain the differences in the amount of compensation they exhibited for some of the stimulus languages.

To conclude, normalization effects were found with a range of different sounds. Firstly, listeners compensated for sentences that were spoken in a second language or even a language that they did not understand (Chapter 4). Secondly, listeners compensated for some types of non-speech sounds (Chapters 2, 3 and 7). Third, Chapter 5 showed that normalization effects can be observed when listeners perform a task that encourages them to focus on pre-categorical properties of the stimuli. These findings show that extrinsic normalization is a relatively general process that operates on a variety of input signals. This shows that normalization effects are caused by a relatively general contrastive process. I also found, however, that the strength of normalization effects can differ among signals. Firstly, despite similar LTAS relations some non-speech signals did elicit normalization whereas others did not (Chapter 2). Secondly, noise contexts induced different effects from speech contexts (Chapter 7). Thirdly, the same precursor sentences had different effects dependent on a listeners background language (Chapter 4). These findings show that central normalization processes must take place at a stage at which at least some amount of specialization has occurred as a result of linguistic exposure. It has

previously been shown that linguistic exposure can influence the encoding of pitch at the level of the brain stem (Krishnan, et al., 2005). This indicates that, in principle, linguistic exposure can influence very early levels of processing. The amount of normalization that a precursor induces on a subsequent target seems to depend on the LTAS relations between a precursor target pair, the spectrotemporal properties of the precursor (and the language-dependent sensitivity that a listener has to those cues), and which hemisphere is most dominantly involved in carrying out the task.

Variance, vowel normalization and other mechanisms

As discussed in the introduction, a listener has access to a number of different mechanisms that work together to allow speech signals to be mapped onto correct phoneme categories. These mechanisms each resolve part of the invariance problem. Interestingly, there seem to be similarities between some of those mechanisms, and the properties of normalization as I found them in this thesis.

Extrinsic normalization influenced vowel perception in an early time-window, namely between 80 and 160 ms after vowel onset. In a recent investigation, Monahan and Idsardi (2010) report findings on *intrinsic* vowel normalization: F_1 was interpreted relative to the F_3 in the same vowel token. F_3 is correlated with vocal tract-length, so the relative perception of F_1 and F_3 helps listeners to deal with vocal-tract dependent differences between speakers. Monahan and Idsardi (2010) found that intrinsic normalization also seems to operate in the N1 time window. While the N1 time window is likely to encompass a number of different processing stages, the similarity in the timing parameters of extrinsic and intrinsic vowel normalization suggests that these different mechanisms both operate at a similar and early processing level.

Reinisch and Sjerps (in prep.) investigated both normalization for speaking rate and extrinsic vowel normalization using eye-tracking. They found that the influence of preceding context, be it durational information for rate normalization or information about the speakers' F_2 for extrinsic vowel normalization, influenced looks to target options on a computer screen no later than actual acoustic differences on the targets did. So a contextually induced difference in perception does not influence subsequent behavior any later than the perceptual effects of actual stimulus differences. This was the case for both compensation for speaking rate and extrinsic vowel normalization. The authors came to the conclusion that these processes must originate from similar, early processing levels.

These findings suggest that a number of compensatory mechanisms – intrinsic normalization, extrinsic normalization and speaking-rate normalization – occur at a similar level, one that is relatively early in the processing hierarchy. It is interesting that while these processes perform computationally very different operations, it seems that they influence perception at more or less the same time. While more research is needed to gain further insight into such similarities, it seems that, at the stage where abstraction processes are leading towards the perception of phonemic representations, a number of mechanisms operate to reduce the influence of input variability.

So, what happens when we listen to another speaker in an everyday conversation? When speech sounds reach the ear of the listener a sequence of processes adjusts perception in a way that is optimized for the situation at hand, in a primarily contrastive manner. These adjustments are similar to the more widely known contrast effects that exist in the visual domain. Extrinsic normalization is one of the core compensation mechanisms that reduces the influence of differences among speakers' vocal tracts. It operates as it were in an orchestra of other mechanisms, including normalization for speaking rate, perceptual learning, audiovisual integration, intrinsic normalization, and compensation for coarticulation. These mechanisms help listeners to perceive speech in a way that makes it feel much less complicated than one might think it is after reading this thesis.

CHAPTER 8: SUMMARY AND CONCLUSIONS

REFERENCES

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America, 116*(5), 3099-3107.

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America, 116*(3), 1729-1738.

Aguilar, L. (1999). Hiatus and diphthong: Acoustic cues and speech situation differences. *Speech Communication, 28*(1), 57-74.

Ainsworth, W. A. (1974). The Influence of Precursive Sequences on the Perception of Synthesized Vowels. *Language and speech, 17*(2), 103 -109.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *In Auditory analysis and perception of speech.* London: Academic Press.

Aravamudhan, R., Lotto, A. J., & Hawks, J. W. (2008). Perceptual context effects of speech and nonspeech sounds: The role of auditory categories. *Journal of the Acoustical Society of America, 124*(3), 1695-1703.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457-474.

Bates, D. M., & Sarkar, D. (2007). lme4: linear mixed-effects models using S4 classes (version 0.999375-27) [software application].

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10-11), 763-786.

Bien, H., Lagemann, L., Dobel, C., & Zwitserlood, P. (2009). Implicit and explicit categorization of speech sounds - dissociating behavioural and neurophysiological data. *European Journal of Neuroscience, 30*(2), 339-346.

Blesser, B. (1972). Speech perception under conditions of spectral transformation .1. Phonetic characteristics. *Journal of Speech and Hearing Research, 15*(1), 5-&.

Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer.

Booij, G. (1995). *The phonology of Dutch.* Oxford: Oxford University Press.

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgements and adaptation level. *Proceedings of the Royal Society of London Series B-Biological Sciences, 151*(944), 384-399.

Broadbent, D. E., Ladefoged, P., & Lawrence, W. (1956). Vowel sounds and perceptual constancy. *Nature, 178*(4537), 815-816.

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System, 36*(1), 22-34.

Cacace, A. T., & McFarland, D. J. (2003). Quantifying signal-to-noise ratio of mismatch negativity in humans. *Neuroscience Letters, 341*(3), 251-255.

Celsis, P., Doyon, B., Boulanouar, K., Pastor, J., Demonet, J. F., & Nespoulous, J. L. (1999). ERP correlates of phoneme perception in speech and sound contexts. *Neuroreport, 10*(7), 1523-1527.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics, 70*(4), 604-618.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America, 118*(3), 1661-1676.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America, 116*, 3668.

Darwin, C. J., McKeown, J. D., & Kirby, D. (1989). Perceptual compensation for transmission channel and speaker effects on vowel quality. *Speech Communication, 8*(3), 221-234.

Diesch, E., Eulitz, C., Hampson, S., & Ross, B. (1996). The neurotopography of vowels as mirrored by evoked magnetic field measurements. *Brain and Language, 53*(2), 143-168.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447-456.

Flege, J. E., Munro, M. J., & Mackay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*(1), 1-26.

Fowler, C. A. (1992). Vowel duration and closure duration in voiced and unvoiced stops - there are no contrast effects here. *Journal of Phonetics, 20*(1), 143-165.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America, 119*(3), 1712-1726.

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience and Biobehavioral Reviews, 25*(4), 355-373.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech, 5*(4), 171-189.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*(3), 363-376.

Gussenhoven, C. (1999). Dutch *Handbook of the International Phonetic Association* (pp. 74–77). Cambridge: Cambridge University Press.

Hansen, T., Walter, S., & Gegenfurtner, K. R. (2007). Effects of spatial and temporal context on color categories and color constancy. *Journal of Vision, 7*(4).

Heinz, J. M., & Stevens, K. N. (1961). On the properties of voiceless fricative constants. *Journal of the Acoustical Society of America*.

Hickok, G., & Poeppel, D. (2007). Opinion - The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Hillenbrand, J. M., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America, 109*, 748.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099-3111.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16*(4), 305-312.

Holt, L. L. (2006a). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America, 120*(5), 2801-2817.

Holt, L. L. (2006b). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America, 119*(6), 4016-4026.

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research, 167*(1-2), 156-169.

Ingram, J. C. L. (2007). *Neurolinguistics: An introduction to spoken language processing and its disorders*: Cambridge Univ Pr.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal-detection-theory and multidimensional-scaling. *Journal of the Acoustical Society of America, 97*(1), 553-562.

Jancke, L., Wustenberg, T., Schulze, K., & Heinze, H. J. (2002). Asymmetric hemodynamic responses of the human auditory cortex to monaural and binaural stimulation. *Hearing Research, 170*(1-2), 166-178.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Proceedings of the first session of the 10th international symposium on spontaneous speech: Data and analysis* (pp. 29 – 54). Tokyo: The National International Institute for Japanese Language.

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

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*(4), 359-384.

Joos, M. (1948). Acoustic Phonetics. *Language, 24*(2), 5-136.

Kiefte, M., & Kluender, K. R. (2005). The relative importance of spectral tilt in monophthongs and diphthongs. *Journal of the Acoustical Society of America, 117*(3), 1395-1404.

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. *Journal of the Acoustical Society of America, 123*(1), 366-376.

Kimura, D. (1961). Cerebral-dominance and the perception of verbal stimuli. *Canadian Journal of Psychology, 15*(3), 166-171.

Kingston, J., & Macmillan, N. A. (1995). Integrality of nasalization and f1 in vowels in isolation and before oral and nasal consonants - a detection-theoretic application of the garner paradigm. *Journal of the Acoustical Society of America, 97*(2), 1261-1285.

Kirk, E. C., & Smith, D. W. (2003). Protection from acoustic trauma is not a primary function of the medial olivocochlear efferent system. *Jaro-Journal of the Association for Research in Otolaryngology, 4*(4), 445-465.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication, 41*(1), 59-69.

Kluender, K. R., & Kiefte, M. J. (2006). Speech perception within a biologically realistic information-theoretic framework. In M. A. Gernsbacher & M. Traxler (Eds.), *Handbook of Psycholinguistics* (2nd ed., pp. 153-199). London: Elsevier.

Krishnan, A., Xu, Y. S., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research, 25*(1), 161-168.

Ladefoged, P. (1989). A note on "Information conveyed by vowels". [Letter]. *Journal of the Acoustical Society of America, 85*(5), 2223-2224.

Ladefoged, P. (1999). American English *Handbook of the International Phonetic Association* (pp. 41–44). Cambridge: Cambridge University Press.

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

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

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

Liegeois-Chauvel, C., Musolino, A., & Chauvel, P. (1991). Localization of the primary auditory area in man. *Brain, 114*, 139-153.

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America, 42*(4), 830-&.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *Journal of the Acoustical Society of America, 102*(2), 1134-1140.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *Journal of the Acoustical Society of America, 113*(1), 53-56.

Loveless, N., Vasama, J. P., Makela, J., & Hari, R. (1994). Human auditory cortical mechanisms of sound lateralization .3. Monaural and binaural shift responses. *Hearing Research, 81*(1-2), 91-99.

Luce, R. D. (1986). *Response times*. New York: Oxford University.

Makela, A. M., Alku, P., & Tiitinen, H. (2003). The auditory N1m reveals the left-hemispheric representation of vowel identity in humans. *Neuroscience Letters, 353*(2), 111-114.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*(5), 407-412.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [S]-[s] distinction. *Perception & Psychophysics, 28*(3), 213-228.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech Language and Hearing Research, 50*(4), 940-967.

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 19-35). Oxford: Oxford University Press.

May, B. J., Budelis, J., & Niparko, J. K. (2004). Behavioral studies of the olivocochlear efferent system - Learning to listen in noise. *Archives of Otolaryngology-Head & Neck Surgery, 130*(5), 660-664.

McCarthy, G., & Donchin, E. (1981). A metric for thought - a comparison of P300 latency and reaction time. *Science, 211*(4477), 77-80.

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

Miller, R. L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America, 25*(1), 114-121.

Mitterer, H. (2006a). Is vowel normalization independent of lexical processing? *Phonetica, 63*(4), 209-229.

Mitterer, H. (2006b). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics, 68*(7), 1227-1240.

Mitterer, H. (2011). Recognizing reduced forms: Different processing mechanisms for similar reductions. *Journal of Phonetics*.

Mitterer, H., Csepe, V., & Blomert, L. (2006). The role of perceptual integration in the recognition of assimilated word forms. *Quarterly Journal of Experimental Psychology, 59*(8), 1395-1424.

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and Cognitive Processes, 25*(6), 808-839.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing* (5th ed.). San Diego: Academic Press.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*(6615), 432-434.

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin, 125*(6), 826-859.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*(5), 2088-2113.

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

Obleser, J., Elbert, T., Lahiri, A., & Eulitz, C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research, 15*(3), 207-213.

Obleser, J., Eulitz, C., & Lahiri, A. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience, 16*(1), 31-39.

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

Pisoni, D. B. (1973). Auditory and phonetic memory codes in discrimination of consonants and vowels. *Perception & Psychophysics, 13*(2), 253-260.

Pisoni, D. B. (1975). Auditory short-term-memory and vowel perception. *Memory & Cognition, 3*(1), 7-18.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Poeppel, D., Phillips, C., Yellin, E., Rowley, H. A., Roberts, T. P. L., & Marantz, A. (1997). Processing of vowels in supratemporal auditory cortex. *Neuroscience Letters, 221*(2-3), 145-148.

Pollack, I., & Pisoni, D. (1971). Comparison between identification and discrimination tests in speech perception. *Psychonomic Science, 24*(6), 299-300.

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social psychology, 18*(1), 10.

R Development Core Team. (2008). R: A language and environment for statistical computing. : Vienna: R Foundation for Statistical Computing.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 37*(3), 978.

Reinisch, E., & Sjerps, M. J. (in prep.). Compensation for speaking rate and spectral voice characteristics take place at a similar point in time.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 40-61.

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

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

Roberts, T. P. L., Flagg, E. J., & Gage, N. M. (2004). Vowel categorization induces departure of M100 latency from acoustic prediction. *Neuroreport, 15*(10), 1679-1682.

Saarinen, J., Paavilainen, P., Schoger, E., Tervaniemi, M., & Näätänen, R. (1992). Representation of abstract attributes of auditory stimuli in the human brain. *Neuroreport, 3*(12), 1149-1151.

Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication, 41*(1), 71-80.

Scott, S. K., & Wise, R. J. S. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition, 92*(1-2), 13-45.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America, 107*(5), 2697-2703.

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics, 68*(1), 1-16.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention , Perception, & Psychophysics, 73*, 1195-1215.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (in press). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention , Perception, & Psychophysics*.

Snyder, E., Hillyard, S. A., & Galambos, R. (1980). Similarities and differences among the P3 waves to detected signals in 3 modalities. *Psychophysiology, 17*(2), 112-122.

Stefanatos, G. A., Joe, W. Q., Aguirre, G. K., Detre, J. A., & Wetmore, G. (2008). Activation of human auditory cortex during speech perception: Effects of monaural, binaural, and dichotic presentation. *Neuropsychologia, 46*(1), 301-315.

Stephens, J. D. W., & Holt, L. L. (2003). Preceding phonetic context affects perception of nonspeech. *Journal of the Acoustical Society of America, 114*(6), 3036-3039.

Stilp, C. E., Alexander, J. M., Kiefte, M. J., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics, 72*, 470-480.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1074-1095.

Summerfield, Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. *Perception & Psychophysics, 35*(3), 203-213.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language, 28*(1), 12-23.

Suzuki, M., Kitano, H., Kitanishi, T., Itou, R., Shiino, A., Nishida, Y., et al. (2002). Cortical and subcortical activation with monaural monosyllabic stimulation by functional MRI. *Hearing Research, 163*, 37-45.

Tavabi, K., Elling, L., Dobel, C., Pantev, C., & Zwitserlood, P. (2009). Effects of Place of Articulation Changes on Auditory Neural Activity: A Magnetoencephalography Study. *PLoS One, 4*(2), e4452.

Tavabi, K., Obleser, J., Dobel, C., & Pantev, C. (2007). Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. *European Journal of Neuroscience, 25*(10), 3155-3162.

Tiitinen, H., Makela, A., Makinen, V., May, P., & Alku, P. (2005). Disentangling the effects of phonation and articulation: Hemispheric asymmetries in the auditory N1m response of the human brain. *BMC Neuroscience, 6*(1), 62.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous Perception and Graded Categorization. Electrophysiological Evidence for a Linear Relationship Between the Acoustic Signal and Perceptual Encoding of Speech. *Psychological Science, 21*(10), 1532-1540.

van Bergem, D. R., Pols, L. C. W., & Beinum, F. J. K.-v. (1988). Perceptual normalization of the vowels of a man and a child in various contexts. *Speech Communication, 7*(1), 1-20.

van Dommelen, W. A. (1999). Auditory accounts of temporal factors in the perception of Norwegian disyllables and speech analogs. *Journal of Phonetics, 27*(1), 107-123.

Van Nierop, D. J. P. J., Pols, L. C. W., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acustica, 29*(2), 110-118.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talkers vowel space. *Journal of the Acoustical Society of America, 60*(1), 198-212.

Wade, T., & Holt, L. L. (2005). Effects of later-occurring nonlinguistic sounds on speech categorization. *Journal of the Acoustical Society of America, 118*(3), 1701-1710.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 90*(6), 2942-2955.

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America, 96*(3), 1263-1282.

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 99*(6), 3749-3757.

Wilson, J. P. (1970). An auditory after-image. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing* (pp. 303-318). Leiden: Sijthoff.

Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csepe, V., et al. (1999). Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cognitive Brain Research, 7*(3), 357-369.

SAMENVATTING EN CONCLUSIES

Samenvatting van de resultaten

Het doel van de serie experimenten die beschreven zijn in dit proefschrift was te onderzoeken hoe luisteraars omgaan met variatie in het spraaksignaal. Wanneer mensen naar spraak luisteren zijn er een aantal cognitieve processen gaande die de luisteraar helpen om met variatie om te gaan. In dit proefschrift onderzocht ik een van deze processen, namelijk compensatie voor spraakkanaalkarakteristieken. Sprekers verschillen onderling in de vorm van hun spraakkanaal. Een van de gevolgen van deze variatie is dat wanneer twee mensen dezelfde klinker uitspreken, bijvoorbeeld /ɛ/, ze zeer verschillende geluiden zullen produceren. In sommige gevallen kunnen deze verschillen zelfs zo groot zijn dat de eerste en tweede formant (F_1 en F_2, deze zijn belangrijk voor het perceptieve onderscheid tussen klinkers) van de [ɛ] van de ene spreker identiek zijn aan die van de [ɪ] van een andere spreker. Luisteraars hebben hier echter nauwelijks last van. Zij kunnen deze verschillende geluiden vrij probleemloos als de juiste foneemcategorieën herkennen. Luisteraars hebben dit vermogen omdat ze voor een deel compenseren voor de algehele spraakkanaalkarakteristieken van de spreker waar ze naar luisteren (Ladefoged & Broadbent, 1957). Ladefoged en Broadbent (1957) lieten zien dat karakteristieken, waargenomen uit een zin, de perceptie van een daaropvolgende klinker beïnvloedden. Deze bevinding vormde de basis voor de experimenten die in dit proefschrift beschreven zijn. In het vervolg van dit hoofdstuk zal ik voor het type effect dat Ladefoged and Broadbent beschreven (dat betekent: de invloed van een zin op de waarneming van een daaropvolgende klinker) de term "extrinsieke normalisatie" gebruiken. Het woord "extrinsiek" refereert daarbij aan het feit dat de waarneming van de klinker wordt beïnvloed door informatie die zich *buiten* de klinker zelf bevindt.

In Hoofdstuk 2 onderzocht ik of extrinsieke normalisatie spraakspecifiek is. Dat wil zeggen, of extrinsieke normalisatie alleen plaatsvindt bij spraaksignalen of bij allerlei signalen. Om dit te onderzoeken manipuleerde ik spraaksignalen op een aantal

manieren, waardoor ik niet-spraaksignalen creëerde. Het eerste experiment had twee beoogde doelen. Ten eerste om het effect van Ladefoged en Broadbent (1957) te repliceren en ten tweede om te onderzoeken of datzelfde effect met niet-spraaksignalen gevonden kon worden. Experiment 1 bestond daarom uit twee delen, een spraak- en een niet-spraakgedeelte. Het spraakgedeelte was een experiment dat werd uitgevoerd in het Nederlands, waarbij proefpersonen woorden categoriseerden die bestonden uit een continuüm van "pit" naar "pet" (dat betekent dat sommige geluiden ambigu waren tussen "pit" en "pet"). Deze woorden werden voorafgegaan door een context zin: "op dat boek staat niet de naam...". Echter, er werden twee versies van deze zin gecreëerd: één waarbij de gemiddelde F_1 relatief laag was en één waarbij de gemiddelde F_1 juist relatief hoog was. De resultaten lieten zien dat proefpersonen de woorden van het continuüm vaker als "pit" waarnamen wanneer het woord voorafgegaan was door de zin met de hoge F_1 dan wanneer deze voorafgegaan was door de zin met de lage F_1 (en vice versa voor "pet"). Hiermee was het eerste doel bereikt: het repliceren van de bevindingen van Ladefoged en Broadbent (1957).

Voor het tweede, niet-spraak, gedeelte van het experiment manipuleerde ik de spraaksignalen, zowel de context zinnen en de doelwoorden (met "doelwoord" refereer ik naar het woord waar de proefpersoon een beslissing over moet nemen), uit het eerste deel door middel van "spectrale rotatie". Spectrale rotatie is een techniek waarbij, als het ware, de informatie in lage frequenties van plaats wisselt met de informatie in hoge frequenties. Deze manipulatie leidt tot signalen die niet op spraak lijken, maar tegelijkertijd zijn de signalen nog steeds even complex als de oorspronkelijke spraaksignalen. Voor dit experiment vroegen we proefpersonen om de spectraal geroteerde versies van het [pɪt] - [pɛt] continuüm te categoriseren. Merk daarbij op dat deze geluiden nu niet langer als "pit" en "pet" klonken maar meer als rare kikkergeluiden (aldus één van mijn proefpersonen). Proefpersonen hadden daarom voorafgaand aan het experiment eerst geleerd om deze geluiden in twee categorieën te verdelen. De resultaten van dit tweede experiment lieten zien dat deze gemanipuleerde materialen dezelfde normalisatie effecten lieten zien als die in het eerste experiment. Hieruit kunnen we dus concluderen dat luisteraars ook normaliseren wanneer ze naar geluiden luisteren die ze niet waarnemen als zijnde spraak.

In een daaropvolgend experiment wilde ik onderzoeken of het normalisatie effect ook gevonden kon worden met stimuli die nog meer manipulaties hadden

ondergaan (zodat ze nog minder overeenkomsten vertoonden met spraak). Hiervoor past ik de volgende manipulaties toe: (1) de materialen werden wederom spectraal geroteerd; (2) De materialen kregen een monotoon toonhoogteverloop; (3) stiltes werden verwijderd (zoals stiltes ten gevolge van sluitingen bij plosieven). Merk daarbij op dat dergelijke stiltes zeer gebruikelijk zijn in spraaksignalen; (4) alle syllaben werden omgedraaid (het effect dat je zou verkrijgen wanneer je, bijvoorbeeld, een grammofoonplaat de verkeerde kant opdraait); (5) Alle syllaben kregen een gelijk geluidsniveau. Na deze manipulaties voerde ik met de nieuwe materialen weer hetzelfde experiment uit zoals beschreven voor Experiment 1. De resultaten van dit nieuwe experiment lieten zien dat er met deze sterk gemanipuleerde materialen *geen* normalisatie effect had plaatsgevonden. Deze bevinding is in strijd met het idee dat normalisatie in principe met alle geluiden plaatsvindt, zoals beschreven in Watkins (1991) en Watkins en Makin (1994). Watkins (1991) heeft namelijk gesuggereerd dat normalisatie effecten het gevolg zijn van puur auditieve processen die niet gevoelig zijn voor de mate waarin de materialen op spraak lijken. Watkins suggereert dat normalisatie effecten met name bepaald worden door de relatie tussen de Gemiddelde Lange Termijn Spectra (GLTS) tussen een doelfoneem (bijvoorbeeld de klinker waar een proefpersoon een beslissing over moet nemen) en de voorafgaande zin. In al mijn experiment waren zulke GLTS relaties steeds zeer vergelijkbaar, en toch vond ik verschillen in het al dan niet optreden van normalisatie.

In een verdere reeks experimenten onderzocht ik of verschillende combinaties van de vijf manipulaties een aspect aan het licht zou brengen dat cruciaal is voor het optreden van normalisatie. De resultaten van deze experimenten lieten zien dat de enige signalen die tot normalisatie effecten leidden in zekere zin tegenovergesteld waren: Experiment 1b (spectraal geroteerde spraak) en Experimenten 4c en 4d (waarop alle manipulaties waren toegepast *behalve* spectrale rotatie). Deze opmerkelijke bevinding laat zien dat er niet een enkel specifiek aspect is dat tot normalisatie leidt. In plaats daarvan lijkt het alsof meer algehele overeenkomsten van niet-spraaksignalen met spraaksignalen belangrijk is voor normalisatie.

Een laatste experiment in dit hoofdstuk onderzocht of de subjectieve aspecten van luisteraars een invloed hadden gehad op de bevindingen. In een nieuw experiment luisterden proefpersonen naar alle "zinnen" (de gemanipuleerde en niet-gemanipuleerde versies) die gebruikt waren in dit hoofdstuk, en moesten ze aangeven hoezeer ze deze geluiden op spraak vonden lijken. Er bleek geen duidelijk verband

tussen de sterkte van normalisatie effecten en de subjectieve oordelen over de gelijkenis met spraak.

In Hoofdstuk 3 onderzocht ik of aandacht een invloed uitoefent op normalisatie. Deze vraag was relevant omdat het mogelijk was dat de verschillen die ik had gevonden tussen de sterkte van het normalisatie effect met verschillende materialen een gevolg zouden kunnen zijn van verschillen in aandacht. Wellicht was het gebrek aan normalisatie effecten met sommige niet-spraak materialen een gevolg van een verminderde aandacht voor de context materialen. Luisteraars dachten wellicht dat de context materialen van de meest extreme niet-spraak materialen om de één of andere reden minder relevant waren voor de waarneming van de daaropvolgende geluiden. Ze zouden daarom geprobeerd kunnen hebben om de context materialen te negeren in de meest extreme niet-spraak experimenten. Om de invloed van aandacht vast te stellen stelde ik een nieuw experiment op, waarin proefpersonen aangespoord werden om ook aandacht aan de context signalen te besteden. Dit gebeurde door middel van een bijkomende taak. Voor deze taak werd de proefpersoon opgedragen om op sommige stimuli geen antwoord te geven. Dit was het geval voor stimuli waarbij de context een plotselinge, sterke afname in amplitude vertoonde. Om deze extra taak goed uit te voeren moesten proefpersonen dus ook aandacht aan de context signalen besteden. Om de sterkte van de invloed van aandacht te meten werden dezelfde materialen gebruikt als in Experimenten 1b en 4c van Hoofdstuk 2. Deze materialen waren namelijk de niet-spraak materialen die in Hoofdstuk 2 een klein normalisatie effect hadden vertoond. Deze materialen zouden daardoor extra gevoelig zijn voor een bijkomend effect van aandacht. De resultaten van dit experiment lieten echter zien dat aandacht niet leidt tot een toename van normalisatie. Dit suggereert dat de sterkte en richting van normalisatie effecten met name afhankelijk zijn van akoestische karakteristieken, en niet van de subjectieve aspecten die proefpersonen er bij ervaren (of het signaal op spraak lijkt of niet).

Hoofdstuk 4 bestond uit een onderzoek dat voortbouwt op een van de conclusies van Hoofdstuk 2. In Hoofdstuk 2 concludeerde ik dat akoestische gelijkenissen met spraak (maar dus niet de subjectieve) tot sterkere extrinsieke normalisatie effecten leidt. Deze bevinding suggereert dat de blootstelling aan taal en spraak die een luisteraar gedurende zijn leven heeft ondergaan een invloed heeft op de mate waarin bepaalde signalen tot extrinsieke normalisatie effecten leiden. Voor Hoofdstuk 4 onderzocht ik daarom of normalisatie effecten verschillen voor sprekers

van verschillende talen. Het experiment dat ik hiervoor opzette was in veel opzichten vergelijkbaar met het eerste experiment van Hoofdstuk 2 (Experiment 1a, waarvoor spraaksignalen gebruikt werden). Dit betekent dat ik een continuüm creëerde van woorden en dat ik context zinnen manipuleerde zodat er een versie met een hoge F_1 en een versie met een lage F_1 ontstonden. Dit deed ik echter voor materialen in het Nederlands, het Engels en het Spaans. Deze materialen legde ik daarna voor aan moedertaalsprekers van het Nederlands, moedertaalsprekers van het Amerikaans-Engels en moedertaalsprekers van het Spaans. Van de moedertaalsprekers van het Spaans werden twee groepen getest: een groep die een beperkte beheersing van het Engels hadden en een groep Spaans-Engels tweetaligen. Al deze luisteraars werd gevraagd woorden van een continuüm te categoriseren. Dit continuüm liep van [sufu] tot [sofo] (de klinkers /u/ en /o/ komen voor in alle drie de talen). De klinkers /u/ en /o/ worden met name onderscheiden door de hoogte van F_1. Deze doelwoorden werden wederom voorafgegaan door zinnen met een relatief hoge of een relatief lage F_1. Deze experimentele opzet maakte het mogelijk om te onderzoeken of de sterkte van extrinsieke normalisatie effecten verschilt tussen mensen met een verschillende taalachtergrond. Dit zou kunnen laten zien of extrinsieke normalisatie afhankelijk is van (1) de mate waarin een luisteraar bekend is met een taal, (2) de status van F_1 als onderscheidende factor in de moedertaal van een luisteraar (voor het Spaans is F_1 mogelijk belangrijker dan voor het Nederlands of Engels), (3) de klinkerinventaris van de moedertaal van een luisteraar, (4) of dat alle verschillen in effectgrootte die optreden tussen materialen van verschillende talen volledig verklaard kunnen worden door GLTS relaties tussen een context en de daaropvolgende klinker.

De resultaten lieten zien dat alle groepen luisteraars compenseren voor spraakkanaalkarakteristieken van de context zinnen (dus: meer /sofo/ antwoorden wanneer de F_1 in de voorafgaande zin laag was dan wanneer deze hoog was). Bovendien vertoonden de proefpersonen ook extrinsieke normalisatie effecten wanneer zij naar een tweede, of volstrekt onbekende taal luisterden. Het complete patroon van de sterkte van normalisatie effecten was echter zeer complex. Spaanse luisteraars vertoonden de meeste normalisatie en dat was het geval voor zinnen in alle drie de talen. Echter, Nederlandse luisteraars vertoonden veel normalisatie met Nederlandse zinnen, iets minder voor Engelse zinnen, en het minst voor Spaanse zinnen. De tweetalige Spaans-Engels sprekers vertoonden een beetje minder normalisatie met Spaanse zinnen dan de Spaans sprekers. De moedertaalsprekers van

het Engels normaliseerden ongeveer evenveel als de moedertaalsprekers van het Spaans. Bovendien lieten extra analyses zien dat de moedertaalsprekers van het Engels verschillen vertoonden in de sterkte van normalisatie over de verschillende materialen. Hetzelfde gold voor de Spaans-Engels tweetaligen.

Hoewel ik niet een enkele simpele verklaring heb kunnen geven voor dit complete patroon zijn er toch een aantal conclusies die uit deze resultaten getrokken kunnen worden. Ten eerste is het belangrijk dat normalisatie effecten gevonden werden ook wanneer proefpersonen naar een tweede of volstrekt onbekende taal luisterden. Dit laat wederom zien dat extrinsieke normalisatie een mechanisme is dat op een brede manier van toepassing is. Ten tweede is een belangrijke bevinding dat verschillen in de sterkte van normalisatie gevonden werden, zelfs tussen verschillende materialen die alleen uit spraaksignalen bestonden. Deze resultaten voegen iets toe aan een van de conclusies van Hoofdstuk 2. Variatie in de sterkte van extrinsieke normalisatie wordt niet alleen gevonden tussen spraak en niet-spraaksignalen, maar ook tussen spraaksignalen onderling, die slechts verschillen in de taal waarin ze geproduceerd zijn. Als gevolg van hun jarenlange blootstelling aan verschillende moedertalen zijn moedertaalsprekers van verschillende talen gevoelig voor andere aspecten van spraaksignalen. De bevindingen in Hoofdstuk 4 laten zien dat verschillen in dergelijke gevoeligheden invloed kunnen hebben op de mate waarin een context zin invloed uitoefent op de waarneming van een daaropvolgend foneem.

Hoofdstuk 5 beschrijft een serie experimenten waarvoor zowel categorisatie- als discriminatie-experimenten werden uitgevoerd om extrinsieke normalisatie te onderzoeken. Voorheen is extrinsieke normalisatie alleen onderzocht door middel van categorisatie experimenten. Categorisatie experimenten leiden er toe dat luisteraars voornamelijk op categorische aspecten van de stimuli letten en minder op informatie die precategorisch is. Een belangrijke stroming in onderzoek naar extrinsieke normalisatie, en context effecten in het algemeen, suggereert dat een belangrijk deel van extrinsieke normalisatie effecten hun oorsprong vinden op precategorische niveaus van verwerking (Kluender & Kiefte, 2006; Sjerps et al., 2011 [Hoofdstuk 2]; Stilp et al., 2010; Watkins, 1991; Watkins & Makin, 1994, 1996). Om deze aanname beter te onderzoeken besloot ik om te onderzoeken of normalisatie effecten ook gevonden kunnen worden in een discriminatie taak. Het wordt aangenomen dat scores die verkregen zijn in een discriminatie taak een betere reflectie zijn van precategorische, auditieve verwerking dan scores in een categorisatie taak (Gerrits &

Schouten, 2004). Voor dit experiment creëerde ik klinker stimuli die een continuüm van [ɪ] naar [ɛ] vormden. Vervolgens creëerde ik twee versies van een [papu] context, een versie waarin de gemiddelde F_1 relatief hoog was en een waarin de gemiddelde F_1 relatief laag was. Daarna plakte ik het continuüm van klinkers voor de [papu] contexten, waardoor voor beide contextversies een continuüm van nonwoorden ontstond van [ɪpapu] naar [ɛpapu]. In een eerste experiment categoriseerden proefpersonen de stimuli van deze continuüms. De resultaten lieten zien dat luisteraars inderdaad een continuüm van [ɪpapu] naar [ɛpapu] waarnamen, maar ook dat hun responsies wederom beïnvloed werden door de F_1 contour in het [papu] gedeelte. Het effect van de context signalen was in dezelfde richting als de effecten die ik in de Hoofdstukken 2, 3 en 4 vond. Dit experiment liet zien dat de materialen in een categorisatie experiment het verwachtte effect konden vertonen.

Vervolgens gebruikte ik dezelfde materialen in een discriminatietaak (4I-oddity). Tijdens deze taak hoorden luisteraars steeds 4 stimuli (een enkele stimulus is bijvoorbeeld [ɛpapu]). In de rij van vier stimuli waren drie stimuli identiek (de "standaard", S) en een stimulus was afwijkend (de "deviant", D). De afwijkende stimulus was altijd ofwel de tweede ofwel de derde stimulus (de volgorde is dus ofwel SDSS, of SSDS). De proefpersonen moesten vervolgens aangeven of de afwijkende stimulus zich in tweede of in derde positie bevond. Hierbij geldt dat wanneer het verschil tussen de deviant en de standaard minder goed te horen is, mensen vaker een verkeerd antwoord zullen geven dan wanneer het verschil zeer duidelijk te horen is. Proefpersonen hoorden als standaard altijd de ambigue stimulus [$^{!}_{\varepsilon}$papu] (waarbij [$^{!}_{\varepsilon}$] aangeeft dat het een ambigue stimulus betreft, die halverwege de Nederlandse klinkers [ɪ] en [ɛ] ligt), en als deviante stimulus altijd [ɪpapu] of [ɛpapu] (beide zijn niet ambigu). Wederom had het [papu] gedeelte ofwel een verhoogde ofwel een verlaagde F_1. Het idee achter dit experiment was dat de invloed van de context op de waarneming van de initiële klinkers er voor zou kunnen zorgen dat, afhankelijk van de context, juist de [ɪpapu] of de [ɛpapu] deviant makkelijker waar te nemen zou zijn. Echter, omdat de discriminatie taak voornamelijk auditieve waarneming weergeeft zou dit alleen het geval zijn als extrinsieke normalisatie plaatsvindt op een precategorisch, auditief niveau van verwerking. De resultaten lieten zien dat het extrinsieke normalisatie effect ook in een discriminatie taak waargenomen kan worden. De discriminatie scores waren inderdaad afhankelijk van de F_1 waardes in de [papu] context signalen. Deze invloed was, wederom, van een contrastieve aard. Een

laatste experiment in dit hoofdstuk bevestigde dat in een discriminatie taak zoals ik in dit hoofdstuk gebruikte, luisteraars zich met name op auditieve aspecten van de stimuli richtten. De resultaten in dit hoofdstuk laten daarmee zien dat perceptuele verschuivingen in klinker waarneming voor een belangrijk deel op een precategorisch verwerkingsniveau plaatsvinden.

Hoofdstuk 6 beschrijft een elektro-encefalografie (EEG) experiment. Dit experiment was opgezet om een beter inzicht te verkrijgen in de temporele ontwikkeling van extrinsieke normalisatie tijdens spraak perceptie. Voor dit onderzoek gebruikte ik stimuli die zeer vergelijkbaar waren met de stimuli die in de experimenten uit het vorige hoofdstuk gebruikt waren. Nu hoorden luisteraars echter een stroom standaard geluiden (wederom het niet bestaande woord [ˈɛpapu]). Deze standaard werd sporadisch afgewisseld door [ɛpapu] of [ɪpapu]. Dit experiment is gebaseerd op de bevinding dat luisteraars een specifieke hersenrespons laten zien wanneer ze een afwijkend geluid horen in een lange rij van dezelfde geluiden. Luisteraars werd gevraagd om op een knop te drukken wanneer zij een afwijkende stimulus hoorden. Er waren twee condities: een waarin het [papu] deel altijd een hoge F_1 had en een conditie waarin het [papu] deel een lage F_1 had. In overeenkomst met het vorige hoofdstuk was de voorspelling dat de hoogte van de F_1 van de [papu] context een invloed zou hebben op de waarneembaarheid van de twee verschillende afwijkende stimuli, waarbij, afhankelijk van de [papu] versie, juist de [ɛpapu] of [ɪpapu] deviant beter waarneembaar zou zijn. Het EEG signaal zou dan kunnen laten zien op welk moment in de tijd zulke verschillen in waarneembaarheid tot stand komen.

In overeenkomst met de resultaten van Hoofdstuk 5 vonden luisteraars het moeilijker om het verschil te horen tussen de standaard [ˈɛ] en de deviant [ɪ] wanneer de F_1 in de [papu] context hoog was in plaats van laag, terwijl het juist moeilijker om het verschil te horen tussen de standaard [ˈɛ] en de deviant [ɛ] wanneer de F_1 in de context laag was in plaats van hoog. Interessanter nog, echter, was het dat dit interactie effect ook zichtbaar was in de EEG signalen die tijdens het experiment waren opgenomen. Uit het EEG signaal bleek dat extrinsieke normalisatie effecten al na ongeveer 120 milliseconden waargenomen konden worden (het N1 tijdsdeel: van 80 tot 160 ms). Dit betekent dat extrinsieke normalisatie al tijdens een relatief vroeg tijdsdeel plaatsvindt. In eerdere tijdsdelen vond ik geen betrouwbare effecten van

normalisatie (tijdens het P1 tijdsdeel: van 30 tot 80 ms). Dit suggereert dat normalisatie geen substantiële invloed heeft vóór het N1 tijdsdeel.

Voor het laatste experimentele hoofdstuk (Hoofdstuk 7) van dit proefschrift onderzocht ik of extrinsieke normalisatie effecten verschillen tussen verwerking in de linker of de rechter hersenhelft. Hiervoor onderzocht ik extrinsieke normalisatie effecten die optreden met spraak en niet-spraak contexten, en voor klinkers die in het linker- of het rechteroor werden afgespeeld. Eerder onderzoek heeft laten zien dat geluiden die in een oor worden afgespeeld in eerste instantie voornamelijk in de tegenovergestelde hersenhelft worden verwerkt (Kimura, 1961; Stefanatos, et al., 2008). Deze bevinding bood mij de kans om te manipuleren in welke hersenhelft de verwerking van een stimulus sterker zou zijn. Wederom gebruikte ik stimuli die vergelijkbaar waren met de stimuli uit de hoofdstukken 5 en 6. Proefpersonen voerden een discriminatie taak uit waarbij ze verschillen moesten waarnemen tussen [ˈɛ] en [ɪ] of tussen [ˈɛ] en [ɛ] in een "4I-oddity" experiment. De [papu] context was wederom gemanipuleerd zodat het een relatief hoge F_1 of een relatief lage F_1 had. Voor de niet-spraak versie creëerde ik twee ruissignalen die dezelfde GLTS hadden als de lage F_1 [papu] versie of de hoge F_1 [papu] versie.

De resultaten lieten zien dat de sterkte van het extrinsieke normalisatie effect sterk afhankelijk was van de experimentele conditie. Over het geheel genomen vond ik dat niet-spraak context signalen een kleiner effect hadden op de waarneming van een daaropvolgende klinker dan spraak contexten (hetgeen in overeenkomst is met de resultaten van Hoofdstuk 2). Echter, ik vond ook dat contrastieve effecten sterker waren wanneer de klinkers in het linkeroor waren aangeboden (welke dus waarschijnlijk sterker in de rechter hersenhelft verwerkt waren) dan wanneer ze in het rechteroor waren aangeboden. In één van de condities vond ik zelfs een integratie-effect (een effect in de omgekeerde richting van de andere context effecten in dit proefschrift). Hoewel dit effect slecht relatief zwak was suggereert het dat verschillen in de sterkte tussen normalisatie effecten wellicht deels een gevolg zijn van elkaar tegenwerkende effecten (zowel integratie als contrastieve effecten). Een dergelijk voorstel zou echter verder onderzocht moeten worden.

Conclusies

Wanneer ik de bevindingen in dit proefschrift samenvoeg komen er een aantal belangrijke aspecten naar voren. Een deel van de bevindingen verschaft inzicht over het verwerkingsniveau waarop extrinsieke normalisatie plaatsvindt. Ten tweede heb ik

informatie verzamelt met betrekking tot mogelijke verschillen in normalisatie tussen de twee hersenhelften. Ten derde heb ik bewijs verzamelt omtrent de verwerkingskarakteristieken van extrinsieke normalisatie, en dan met name de vraag of er bepaalde voorwaarden zijn waaraan een signaal moet voldoen om extrinsieke normalisatie effecten tot gevolg te hebben. Deze drie aspecten zal ik in de nu volgende secties apart bespreken. In de laatste sectie zal ik bespreken hoe extrinsieke normalisatie effecten, in samenwerking met andere mechanismen, de luisteraar helpen om te gaan met variabiliteit in het spraaksignaal.

De plaats van verwerking in de cognitieve hiërarchie.

Eén van de bijdragen van dit onderzoek aan de bestaande literatuur heeft betrekking op het verwerkingsniveau waarop extrinsieke normalisatie plaatsvindt. Om dit aspect te bespreken moet echter eerst een verschil worden gemaakt tussen het normalisatieproces dat ik heb onderzocht en een ander proces dat functioneel verschilt van extrinsieke normalisatie, maar dat vergelijkbare invloeden kan hebben op waarneming. In de experimenten die beschreven zijn in dit proefschrift heb ik geprobeerd om "centrale" in plaats van "perifere" context effecten te onderzoeken. Om uit te leggen wat ik met perifere effecten bedoel zal ik een voorbeeld geven. Stel, men creëert een geluid van ongeveer één seconde, waarvan het spectrum dalen bevat op dezelfde frequenties waar het spectrum van een klinker pieken bevat (met andere woorden, het complement van die klinker). Als dit geluid wordt afgespeeld met daaropvolgend een ruissignaal met een uniform spectrum, dan horen luisteraars de complementaire klinker in het ruis signaal. Echter, dit effect, wat ook wel het "auditory after image" effect genoemd is, verdwijnt wanneer het eerste signaal korter is dan 150 ms, wanneer er een stilte van 500 ms of meer tussen de context en het tweede signaal zit, of wanneer de context en het tweede signaal worden afgespeeld in een verschillend oor (Summerfield, et al., 1984). Deze bevindingen laten zien dat er een type effecten bestaat dat slecht van korte duur is, en dat geen invloed heeft op de perceptie van signalen die in een ander oor zijn afgespeeld. Dit type effecten worden "perifere effecten" genoemd. In dit proefschrift heb ik mijn aandacht gericht op "centrale effecten". Centrale effecten worden waargenomen over langere intervallen tussen context en doelsignaal, en ook bij contralaterale presentatie (zie Hoofdstuk 2 tot en met 7, en bijvoorbeeld Watkins, 1991). De effecten die beschreven zijn door Watkins (1991), en de effecten in dit proefschrift zijn dus functioneel gezien verschillend dan de effecten die beschreven zijn door Summerfield, et al. (1984). De

veronderstelde dissociatie tussen perifere en centrale effecten werd versterkt door de bevindingen zoals beschreven in Hoofdstuk 6. Ik vond namelijk geen contrastieve invloeden tijdens het P1 tijdsdeel (van 30 - 80 ms), terwijl twee controle experimenten lieten zien dat het experimentele ontwerp in principe wel effecten aan het licht zou kunnen brengen in dit tijdsdeel. Dit suggereert dat extrinsieke normalisatie nog geen invloed uitoefent voor ongeveer 80 ms na het begin van het klinker signaal.

De resultaten van Hoofdstuk 6 lieten zien dat er wel extrinsieke normalisatie effecten optraden in het tijdsdeel van 80 tot 160 ms. De dominante component in dat tijdsdeel is de N1 component. Deze component is beschreven als zijnde een gevolg van processen die plaatsvinden op de grens tussen auditieve verwerking en fonemische verwerkingsniveaus (Näätänen & Winkler, 1999; Roberts, et al., 2004; Tavabi, et al., 2007). Het is daarom aannemelijk dat extrinsieke normalisatie perceptie beïnvloedt op of net voor het moment dat ook de belangrijkste aanwijzingen voor foneemidentiteit aan het signaal worden onttrokken. Deze conclusie wordt ondersteund door de bevindingen in Hoofdstuk 5, waarin normalisatie effecten werden gevonden in een taak waarbij luisteraars werden aangespoord om zich met name te richten op precategorische aspecten van de stimuli.

Gelateraliseerde processen?

In Hoofdstuk 7 onderzocht ik lateralisatie van contextuele invloeden op perceptie. De resultaten lieten zien dat wanneer een doelklinker in het linkeroor werd afgespeeld de context signalen een contrastieve invloed hadden. Dit was het geval wanneer de context uit spraak bestond, maar ook als het uit ruis bestond. Wanneer een doelklinker in het rechteroor werd afgespeeld vond ik alleen een contrastief context effect wanneer de context uit spraak bestond. In de Hoofdstukken 2 en 4 suggereerde ik dat blootstelling aan taal gevolgen kan hebben gehad voor de invloed die een voorafgaand signaal heeft op de perceptie van een doelklinker. De bevindingen uit Hoofdstuk 7 suggereren dat de invloed van blootstelling aan taal met name context effecten in de linker hersenhelft beïnvloedt. Onder normale omstandigheden komen spraaksignalen via beide oren binnen. De grootte van het extrinsieke normalisatie effect onder dergelijke omstandigheden is dan waarschijnlijk een combinatie van de afzonderlijke invloeden van de twee hersenhelften, en de mate waarin een specifieke taak in meerdere of mindere mate op verwerking in een van de hersenhelften berust.

De bevindingen in Hoofdstuk 7 hebben betrekking op een recente theorie over taal en spraak verwerking. Deze theorie, de zogenaamde "Asymmetric Sampling in

Time" (AST) hypothese (Poeppel, 2003), stelt dat verschillen in de verwerking tussen de hersenhelften optreden omdat de twee hersenhelften informatie integreren over verschillende tijdsspannen. De rechter hersenhelft zou informatie integreren over langere tijdspannes dan de linker hersenhelft. Als dit het geval zou zijn dan zou men verwachten dat in de rechter hersenhelft integratieve invloeden van context zouden optreden, omdat integratie over langere tijdsdelen er toe zouden moeten leiden dat context en de doelklinker als een object worden waargenomen. Voor de linker hersenhelft, die over kortere tijdsspannen zou integreren, worden context en doelklinker dan juist eerder als aparte objecten waargenomen. Daarbij zouden dan contrastieve effecten moeten optreden. De resultaten in Hoofdstuk 7 laten, paradoxaal genoeg, een effect in precies de tegenovergestelde richting zien van datgene wat men op basis van AST zou verwachten. De sterkste contrastieve effecten werden gevonden voor doelklinkers waarvan de dominante verwerking in de rechterhersenhelft had plaatsgevonden. Een bijkomende bevinding was dat er in één conditie een integratief effect optrad. Dat was het geval wanneer de doelklinker in het rechteroor was afgespeeld en een ruis context in het linkeroor. Dit enkele geval suggereert dat het ontbreken van een contrastief effect met ruissignalen wanneer de doelklinker in het rechteroor wordt afgespeeld het gevolg zou kunnen zijn van elkaar tegenwerkende krachten: contrastieve en integratieve. Deze suggestie biedt een mogelijke verklaring voor een aantal studies die in het verleden nogal variabele normalisatie effecten hebben laten zien (en soms zelf integratieve effecten) met niet-spraaksignalen (Aravamudhan, et al., 2008; Fowler, 1992; Mitterer, 2006b; van Dommelen, 1999). Een dergelijk voorstel zou echter verder onderzocht moeten worden. Ten eerste is het belangrijk om op een directere manier te onderzoeken in welke mate de manipulatie die ik gebruikt heb voor lateralisatie (het oor waarin stimuli gepresenteerd worden) ook in deze experimentele opzet voor lateralisatie in de verwerking zorgt. Ten tweede is het belangrijk om het integratieve effect te repliceren omdat het in slechts één conditie werd gevonden.

Voorwaarden voor extrinsieke normalisatie: spraak specifiek?

In dit proefschrift heb ik een aantal mogelijke "voorwaarden" voor extrinsieke normalisatie onderzocht. Extrinsieke normalisatie effecten werden waargenomen met een variatie aan verschillende soorten stimuli. Ten eerste normaliseerden luisteraars voor zinnen die uitgesproken waren in een tweede of zelfs compleet onbekende taal (Hoofdstuk 4). Ten tweede normaliseerden luisteraars voor sommige typen niet-

spraak geluiden (Hoofdstukken 2, 3 en 7). Ten derde lieten de experimenten uit Hoofdstuk 5 zien dat normalisatie effecten ook optreden wanneer luisteraars een taak uitvoeren waarbij ze zich sterker op auditieve aspecten van de stimuli richten. Deze bevindingen laten zien dat extrinsieke normalisatie processen vrij algemene processen zijn die op veel verschillende signalen van toepassing zijn. Dit suggereert dat normalisatie effecten het gevolg zijn van een relatief vroeg, met name contrastief proces. Ik vond echter ook dat de sterkte van het normalisatie effect kan verschillen tussen signalen. Ten eerste vond ik dat sommige niet-spraaksignalen wel een effect en anderen weer geen normalisatie effect vertoonden ondanks het feit dat de GLTS relaties tussen context en doelgeluid vergelijkbaar waren (Hoofdstuk 2). Ten tweede leidden ruiscontexten tot andere effecten dan spraak contexten (Hoofdstuk 7). Ten derde verschilden de effectgroottes van dezelfde stimuli afhankelijk van de moedertaal van de proefpersonen. Deze bevindingen laten zien dat centrale normalisatie effecten op een niveau plaatsvinden waar in elk geval enige taalspecifieke specialisatie in de verwerking heeft kunnen plaatsvinden. De twee belangrijkste conclusies zijn daarom dat normalisatie effecten vrij vroeg en algemeen zijn, maar dat ze op een niveau plaatsvinden waar enige taalspecifieke specialisatie heeft plaatsgevonden. Deze twee conclusies lijken op het eerste gezicht vrij tegenstrijdig, maar recent onderzoek heeft laten zien dat taalspecifieke invloeden al op relatief lage niveaus plaatsvinden. Zo blijkt dat taalachtergrond de verwerking van toonhoogte informatie beïnvloedt ter hoogte van de hersenstam (Krishnan, et al., 2005). Dit laat zien dat, in principe, blootstelling aan spraak in een specifieke taal zeer vroege verwerkingsniveaus kan beïnvloeden. De sterkte van normalisatie effecten die een bepaalde context stimulus heeft op en daaropvolgend geluid lijkt daarom afhankelijk van de GLTS relaties tussen een context en het doelgeluid, de spectrotemporale opbouw van de context (en dan met name de taalafhankelijke gevoeligheid van de luisteraar voor deze aspecten), en de hersenhelft die het sterkst betrokken is bij de uitvoering van een bepaalde taak.

Variatie, extrinsieke normalisatie, en andere mechanismen

Zoals besproken in de introductie van dit proefschrift heeft de luisteraar beschikking over een aantal verschillende cognitieve mechanismen. Deze mechanismen werken samen om de luisteraar te helpen bij het interpreteren van geluiden als behorende tot een bepaalde foneemcategorie. Elk van deze afzonderlijke mechanismen lossen een deel van het variatie probleem op. Het blijkt dat er

interessante overeenkomsten te vinden zijn tussen sommige van deze mechanismen en het normalisatie mechanisme dat ik hier onderzocht heb.

Ik vond dat extrinsieke normalisatie de perceptie van klinkers beïnvloedde in een vroeg tijdsdeel, namelijk van 80 tot 160 ms (N1) na het begin van de klinker. In een recent artikel rapporteren Monahan and Idsardi (2010) iets vergelijkbaars over hun bevindingen met *intrinsieke* klinkernormalisatie. Intrinsieke normalisatie berust op het feit dat F_1 relatief wordt waargenomen aan F_3 binnen eenzelfde klinker. F_3 correleert met de lengte van het spraakkanaal, dus de relatieve perceptie van F_1 aan F_3 helpt luisteraars om om te gaan met verschillen tussen sprekers die te verklaren zijn op basis van de lengte van het spraakkanaal. Monahan and Idsardi (2010) vonden dat intrinsieke normalisatie ook opereert tijdens het N1 tijdsdeel. Het moet gezegd dat er zich meerdere verwerkingsniveaus binnen het N1 tijdsdeel kunnen afspelen, maar het is interessant om te zien dat er overeenkomsten zijn tussen de timing van zowel extrinsieke als intrinsieke normalisatie processen, en dat de timing suggereert dat het om relatief vroege processen gaat.

Verder hebben Reinisch and Sjerps (in prep.) recent onderzoek gedaan naar extrinsieke klinkernormalisatie en naar duurnormalisatie met behulp van oogbewegingsonderzoek. Wij onderzochten daarvoor het Nederlandse "a" "aa" contrast, dat zowel op duur als op de hoogte van F_2 berust (onder andere). Daarbij vonden wij dat de invloed van voorafgaande context, voor zowel duur informatie als informatie over de gemiddelde F_2 voor extrinsieke normalisatie, oogbewegingen naar verschillende woorden op een scherm beïnvloedden (bijvoorbeeld "gas" vs. "gaas"). De invloed van normalisatie (dus, voorafgaande context) op oogbewegingen werd op hetzelfde moment zichtbaar als wanneer er daadwerkelijk akoestische verschillen zaten tussen gehoorde doelklinkers. Dit laat zien dat perceptuele verschillen die het gevolg zijn van context zich in gedrag niet later manifesteren dan daadwerkelijke akoestische verschillen. Dit was het geval voor extrinsieke klinkernormalisatie en voor duurnormalisatie. Reinisch and Sjerps redeneerden daarom dat deze twee processen beide op een vroeg verwerkingsniveau plaatsvinden.

Deze gecombineerde bevindingen suggereren dat een aantal compensatiemechanismen - intrinsieke normalisatie, extrinsieke normalisatie, en duurnormalisatie - op een vergelijkbaar, vroeg, verwerkingsniveau plaatsvinden. Het is interessant dat dergelijke overeenkomsten gevonden worden ondanks het feit dat deze processen computationeel gezien zeer verschillende problemen oplossen. Het is

duidelijk dat meer onderzoek naar dergelijke overeenkomsten nodig is, maar het lijkt er op dat op het verwerkingsniveau waarop abstraherende processen meer en meer richting foneemperceptie leiden, er een aantal mechanismen in werking treden die de invloed van variabiliteit verminderen.

Dus wat gebeurt er nou precies wanneer wij in een alledaagse situatie naar spraak luisteren? Wanneer spraakgeluiden het oor van een luisteraar bereiken treedt er een reeks van mechanismen in werking die perceptie trachten te optimaliseren aan de huidige situatie, op een hoofdzakelijk contrastieve manier. De aanpassingen in perceptie die één gevolg zijn van deze processen vertonen overeenkomsten met de welbekende contrast effecten in visuele verwerking. Extrinsieke normalisatie is een van de belangrijkste compensatiemechanismen die variabiliteit als gevolg van verschillen in spraakkanaal lengte verminderen. Dit proces voltrekt zich in een samenspel met een hele reeks andere mechanismen zoals duurnormalisatie, perceptueel leren, audiovisuele integratie, intrinsieke normalisatie en compensatie voor coarticulatie. Al deze mechanismen zorgen er voor dat luisteraars spraak waarnemen op een manier die een stuk minder gecompliceerd voelt dan men zou denken na het lezen van dit proefschrift.

SAMENVATTING EN CONCLUSIES

CURRICULUM VITAE

Matthias J. Sjerps was born in 1982 in Amsterdam, The Netherlands. He obtained a Bachelor's degree in Linguistics, with specializations in Psychology and Phonetics, from the University of Utrecht. He obtained a Master's degree in Cognitive Neuroscience, with a specialization in Psycholinguistics, from the Radboud University of Nijmegen. In 2007 he was awarded a 3-year scholarship from the Max Planck Society to do his PhD research at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. He joined the Language Comprehension Group. In 2010 he was awarded a grant to spend four months on a research project at the University of Texas at Austin, USA. Currently he is working as postdoctoral researcher in the Individual Differences Group at the Max Planck Institute for Psycholinguistics.

CURRICULUM VITAE

MPI Series in Psycholinguistics

---

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

21. Fixed expressions and the production of idioms. *Simone A. Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*
35. Retrieving semantic and syntactic word properties. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*
37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*
42. The acquisition of auditory categories. *Martijn Goudbeek*

43. Affix reduction in spoken Dutch. *Mark Pluymaekers*
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*
45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*
52. Prosodic structure in speech production and perception. *Claudia Kuzla*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*
54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*
55. Processing the fine temporal structure of spoken words. *Eva Reinisch*
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*

64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*