Matthias Franken

LISTENING FOR SPEAKING

Investigations of the relationship between speech perception and production

DONDERS SERIES

Listening for speaking

Investigations of the relationship between speech perception and production

Matthias K. Franken

Listening for speaking: Investigations of the relationship between speech perception and production
PhD thesis, Radboud University Nijmegen

ISBN 978-94-6284-131-4
Cover artwork by Franke Hoogerkamp - www.frankemargrete.nl
Printed by Ipskamp printing

This research was carried out at the Donders Graduate School for Cognitive Neuroscience and the International Max Planck Research School for the Language Sciences. The research was funded by a research talent grant from the Netherlands Organisation for Scientific Research (NWO) and the Donders Graduate School for Cognitive Neuroscience

Listening for speaking

Investigations of the relationship between speech perception and production

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,

volgens besluit van het college van decanen

in het openbaar te verdedigen op maandag 5 februari 2018

om 12.30 uur precies

door

Matthias Karl Marleen Franken

geboren op 7 september 1989

te Antwerpen, België

Promotoren:
Prof. dr. Peter Hagoort
Prof. dr. James M. McQueen

Copromotoren:
Dr. Daniel J. Acheson (Uptake, Chicago, USA)
Dr. Jan-Mathijs Schoffelen

Manuscriptcommissie:
Prof. dr. Antje Meyer
Prof. dr. Niels O. Schiller (Universiteit Leiden)
Prof. dr. Robert J. Hartsuiker (Universiteit Gent, België)

TABLE OF CONTENTS

As a Flemish PhD candidate living in the Netherlands, I have experienced some of the intricacies of speech production and speech perception first-hand. After having lived for a few years in an environment where Dutch is spoken in an accent quite unlike my native accent, I find myself inadvertently adopting some typically Dutch phonetic and linguistic features in my own speech. This illustrates just one of the many ways in which speech perception and speech production interact. This thesis is dedicated to all those who have joined me, over the years, in laughing, marveling and wondering about some of these intriguing phenomena.

1

INTRODUCTION

1.1 SPEECH PERCEPTION AND PRODUCTION

Speech is an amazing phenomenon. It allows us to translate the most complex stories, ideas and concepts into rich acoustic sequences, in such a way that listeners are able to reconstruct the original message. This remarkable skill requires us to control about 100 muscles spread across the vocal tract, tightly coordinating and synchronizing them, producing around 5-10 syllables per second (Pellegrino, Coupé, & Marsico, 2011). If we stop and think about this, producing speech seems like a very daunting task. Yet somehow, almost every human infant manages to acquire this skill effortlessly in just a few years.

In addition to producing speech, we also manage to perceive and comprehend speech produced by others. This task is no less daunting. Starting from a noisy, highly variable acoustic sequence, we have to reconstruct the speaker's original message. One of the most difficult aspects of speech perception is indeed the high variability of the acoustic input. Not only do speakers vary in accents, dialects, languages, voice characteristics, etc., but a single speaker's speech may also vary from time to time depending on various linguistic factors (coarticulation, prosody, …) as well as non-linguistic factors (mood, emotion, social context, etc.). Despite this variation, we are still somehow able to map this noisy acoustic input onto appropriate linguistic representations almost without effort.

A long-standing debate in the speech sciences is concerned with the interaction between the systems that carry out these tasks, speech production and perception. While production and perception have long been investigated separately, researchers have started to investigate their interaction in more recent years. These lines of research have revealed complex interactions between speech production and perception, and show that it is important to not only study perception or production in isolation. Ultimately, speakers produce speech in order for a listener to understand or reconstruct the original message through their speech perception system, and listeners try to reconstruct this message by processing a signal that was generated by the speaker's production system. In addition, in everyday conversational settings, people often very quickly switch roles from being the speaker to being a listener and vice versa. Therefore, it is not surprising that many researchers have proposed a tight link between speech perception and production systems, with some even arguing that speech perception crucially involves recruitment of the motor system (Liberman & Mattingly, 1985; Möttönen & Watkins, 2011; Pickering & Garrod, 2007).

Previous research on the perception-production relationship in speech has shown complex interactions, based on studies showing correlations between the two domains, as well as studies showing directed influences (speech perception affecting speech production and vice versa). First, correlational studies have shown associations between idiosyncrasies in perception and production (Ghosh et al., 2010; Perkell, Guenther, et al., 2004; Perkell, Matthies, et al., 2004). The overall idea is that if perception and production are tightly linked, over time, individual variability in perception or production should lead to co-variability in the other domain. For example, these studies have suggested that differences in auditory discrimination abilities are associated with differences in production variability. Other studies have associated auditory prototypes with speech production (Newman, 2003). Newman showed that listeners' perceptual prototypes for stop consonants corresponded to the listeners' own articulation of those consonants. In contrast, additional studies have shown that in some cases associations between perception and production could not be found or were very small indeed (Kraljic, Brennan, & Samuel, 2008). As Beddor (2015) notes, this is to be expected given that speech perception is known to be much more malleable than speech production. Perception needs to be highly flexible to accommodate the large variability in the speech signal. The lack of a clear direct link with production becomes evident in cases where listeners are able to understand accented speech they cannot reproduce themselves.

Second, beside simple associations between idiosyncrasies in perception and production, investigators have examined direct influences of one domain on the other. For example, a broad range of studies have looked into whether and how perception might influence production. The clearest case of perceptual influence on production may be what is known as phonetic convergence (Pardo, 2006). This is the phenomenon where speakers tend to inadvertently mimic phonetic details produced by their interlocutor. Although it may be tempting to explain such convergence as a fast, automatic priming mechanism from perceived speech to subsequently produced speech, studies have shown that the link is not so straightforward (Pardo, 2012). In fact, studies have shown both convergence and divergence, depending on a broad variety of social and pragmatic factors. In addition, another line of research has focused on the inception and spread of sound changes. In particular, ongoing sound changes provide the opportunity of investigating associations between perception and production of the sound change in question (Beddor, 2015). For example, Harrington, Reubold & Kleber (2008) report a study on /u/-fronting, an ongoing sound change in standard southern British English. The results indicate that listeners who produce the innovative variant (with /u/-fronting), also show evidence of a shifted perceptual boundary. This suggests that perceptual phonological boundaries vary across individual listeners and are strongly related to differences in speech production in the same individuals. In another study on the same sound change (Kleber, Harrington, & Reubold, 2012), the same authors show that in this case a perceptual change precedes the production of the innovative (fronted) variant, suggesting that at an intermediate stage, individuals will show a perceptual change without producing the innovative variant. Yet another case of speech perception affecting speech production is concerned with perception of one's own speech during speech production, or auditory feedback. In this context, the perception of one's own voice may be used in a more direct way to control, monitor, maintain and/or update one's speech productions. As the main topic of this thesis is to investigate the role of auditory feedback during speech production, the literature on this topic will be reviewed in more detail below.

Finally, some investigations have focused on how speech production may affect speech perception. Initially, this may seem odd, as most researchers assume that perception crucially drives production. For example, it is commonly assumed that perception precedes production in development and learning (Escudero, 2007; Kuhl et al., 2008), as infants and language learners would need to perceive a phonological contrast in order to be able to produce it themselves. However, some studies have shown evidence of influences from production on perception. Studies using altered auditory feedback suggest that speech motor adaptation affects speech perception (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014; Schuerman, Nagarajan, McQueen, & Houde, 2017). Another study, looking at language learners, showed that production of the to-be-learned sounds could disrupt formation of perceptual representations (Baese-Berk & Samuel, 2016).

Overall, the research summarized above suggests a complex interplay between speech perception and speech production. This complex interaction shows that in order to fully understand how speech production and speech perception work, it is vital to study how they interact. While sensorimotor interactions have been studied to a great deal in other motor systems like reaching or eye movements (Shadmehr & Mussa-Ivaldi, 2012; Wolpert & Ghahramani, 2000), studying sensorimotor integration in speech is particularly important given the much more complex nature of the speech signal and the speech motor system. The current thesis reports investigations of this perception-production interplay, with a particular emphasis on the role of auditory feedback in speech production.

1.2 AUDITORY FEEDBACK

Interestingly, speech production almost never occurs in isolation. While one can imagine someone listening without speaking, speaking almost always includes perceiving the results of one's own articulations. The perceived signal is commonly referred to as auditory feedback.

Already early on, researchers noticed that the perception of one's own voice is not a trivial by-product of speech production, but instead that one's voice is actually being monitored by the speech production system. Numerous studies have shown that delaying auditory feedback by about 200ms leads to speech errors (Fairbanks & Guttman, 1958; Lee, 1950). This suggests that although online auditory feedback may not be strictly necessary for speech production, disturbed auditory feedback does affect speech. So for what purpose does the production system use auditory feedback? Research in speech motor control and motor control in general has proposed two main functions of feedback. Sensory feedback may be used to (1) distinguish sensations that are self-generated from those generated by others (Eliades & Wang, 2008; Schütz-Bosbach, Mancini, Aglioti, & Haggard, 2006), and thus lead to a sense of agency (Hickok, 2012; Levelt, Roelofs, & Meyer, 1999; Lind, Hall, Breidegard, Balkenius, & Johansson, 2014), and (2) to monitor our motor performance and errors (Hartsuiker & Kolk, 2001; Levelt et al., 1999). It is the latter function of auditory feedback that is the topic of the current thesis.

Several lines of research in recent decades have focused on subtler alterations of auditory feedback than a simple delay. The general idea of these studies is to record participants' speech, manipulate it, and play it back to them in real-time (i.e., without noticeable delays), effectively mimicking a speech error. Broadly, studies using this technique of altered auditory feedback use one of two paradigms (looking either at immediate responses to unexpected feedback perturbations (Burnett, Freedland, Larson, & Hain, 1998; Elman, 1981; Liu & Larson, 2007), or looking at longer-term adaptations in response to consistently altered auditory feedback (Houde & Jordan, 1998; Jones & Munhall, 2000; Purcell & Munhall, 2006a; Villacorta, Perkell, & Guenther, 2007). Manipulations that have been applied include pitch shifts, formant shifts, loudness modulations (Bauer, Mittal, Larson, & Hain, 2006; Lane & Tranel, 1971) and fricative noise manipulations (Casserly, 2011; Shiller, Sato, Gracco, & Baum, 2009).

In the first type of paradigm, unexpected feedback perturbations are applied to investigate the immediate response to the perturbation. Most of the studies that employ this paradigm have used pitch shifts, but formant manipulations have also been used (Purcell & Munhall, 2006b; Tourville, Reilly, & Guenther, 2008). Broadly speaking, about 100-150ms after the perturbation kicks in, speakers tend to respond by changing their speech output in the opposite direction of the perturbation, essentially compensating for the feedback they are receiving. For example, when speakers hear themselves at a higher pitch than expected, they respond by lowering the pitch in their speech output (Burnett et al., 1998). According to the dominant view on auditory feedback in speech motor control, a comparison between the expected (i.e., predicted) auditory feedback and the perceived auditory feedback yields a prediction error, which in turn leads to generating compensatory motor commands. This way, the production system tries to minimize the prediction error. The behavioral responses happen not only relatively quickly, but are also automated and therefore not under conscious control (Hain et al., 2000).

The second paradigm focuses on adaptations of the articulatory commands as a consequence of continued exposure to altered feedback. In this paradigm, speakers are exposed to altered feedback over the course of multiple trials. Usually, feedback returns back to normal afterwards in a wash-out phase. The main finding in this paradigm is that speakers show a change in speech production that compensates (partially) for the alteration, even after feedback has returned to normal (Jones & Munhall, 2000; Purcell & Munhall, 2006a). This suggests that speakers have adapted their articulatory commands as a result of the altered auditory feedback. Overall, these lines of research show that auditory feedback is not only monitored for online error correction, but also for maintaining and/or updating feedforward speech motor control.

Curiously, studies on auditory feedback have not only reported that speakers compensate or adapt by shifting their speech in the opposite direction compared to the feedback perturbation, but some studies have reported following responses: speakers shifting their speech in the direction of the feedback perturbation (Behroozmand, Korzyukov, Sattler, & Larson, 2012; Burnett et al., 1998). This is unexpected: it is unclear why speakers, when hearing themselves speak at a higher pitch than expected, would increase their pitch even more. The dominant view on auditory feedback in speech production cannot explain this, as speakers should try to minimize the discrepancy between expected and observed auditory feedback. Although most studies have actually not paid much attention to following responses, merely reporting them if mentioning them at all, Hain et al. (2000) attempt to explain following responses by the perceived source of the feedback. That is, if the speaker considers the auditory input as being self-generated, an opposing response is generated, while if the auditory signal is considered to be generated by an external source, a following response is made. Consider the case of a choir singer: if they hear themselves at a lower-than-expected pitch, they should increase pitch so as to not sound too flat. On the other hand, if they hear their fellow singers at a lower-than-expected pitch, they should decrease pitch to match the pitch of the other choir members to sing in tune.

1.3 MODELING FEEDBACK IN MOTOR CONTROL

Recent research and literature on auditory feedback in speech motor control has come to align with the broader (non-speech) motor control literature. For example, studies in visuomotor control have focused on the role of sensory feedback in arm reaching. Similar to the acoustic feedback manipulations described above, robotic manipulations of visual and/or somatosensory feedback during reaching tasks have shown that people compensate for perturbations and show adaptive aftereffects when feedback is restored (Shadmehr & Mussa-Ivaldi, 1994). Dominant models in sensorimotor control suggest the use of a so-called internal forward model (Wolpert, Ghahramani, & Jordan, 1995; Wolpert & Ghahramani, 2000). The forward model is an internal neural model of the motor plan, which makes it possible to predict the sensory consequences of issued motor commands. In fact, when motor commands are generated and sent to the muscles, a copy (the "efference copy") is used by the forward model to predict what the sensory consequences of the action will be. Such prediction allows for comparison with the observed sensory feedback and fast error monitoring. Within forward models of sensorimotor control, sensory feedback can be used for issuing corrective motor commands when necessary, as well as for learning, maintaining, and updating the internal forward models.

The findings and modeling work in general motor control have led to applications of these ideas to the speech control domain (Figure 1.1). While the results and ideas generated by studying arm reaching movements are generally hypothesized to apply to motor control in general, applications of these ideas in the speech have been challenging due to the complexity of speech motor control. While the arm is a fairly simple motor system with a limited set of muscles, producing speech involves controlling and coordinating many more muscles spread across the vocal tract with many more degrees of freedom than the relatively simple muscles in the arm. In addition, while most visuomotor research deals with visual feedback about the arm's or hand's position (i.e., in 3D space), auditory feedback involves a more complex set of dimensions, including pitch, several formants, loudness, as well as complex spectrotemporal integrations of these parameters. Therefore, it is worth investigating to what extent the ideas based on a fairly simple motor control model will apply to this more complex system.

One of the dominant models in speech motor control is the DIVA model (Directions Into Velocities of Articulators) (Guenther, Ghosh, & Tourville, 2006; Guenther & Vladusich, 2012; Guenther, 2016; Tourville & Guenther, 2011). In brief, the DIVA model consists of a feedforward control system and a feedback control system (where the latter can be subdivided in an auditory control subsystem and a somatosensory control subsystem). The feedforward control system allows for stored motor programs to be executed (black arrows in Figure 1.1), while the feedback control systems compare the sensory targets, or expected sensory consequences of articulations, with the observed sensory input (auditory and somatosensory feedback, see red and blue arrows in Figure 1.1). In case of mismatches, the prediction errors lead to the generation of compensatory motor commands (e.g., responses to unexpected feedback, green arrow in Figure 1.1). The final motor commands sent to the articulators are a summation of motor commands generated by the feedforward and feedback control systems. In addition, the compensatory motor commands generated by the feedback control system can be used to update or adapt the feedforward control system, thus adapting to sustained altered feedback to avoid future errors. Another dominant model of speech motor control is the state feedback control model (SFC) of speech (Houde, Kort, Niziolek, Chang, & Nagarajan, 2013; Houde & Nagarajan, 2011), which constitutes a more direct application of ideas from non-speech motor control to the speech domain. The critical controlled variable in the SFC model is an internal representation of the state of the speech production system. As the actual state is not accessible, it is estimated based on the past motor commands (via the efference copy), and it can be updated by prediction errors resulting from the detection of any feedback mismatches. Additional speech production models have been developed, among others extending these ideas to speech sequencing, to higher cognitive levels (Hickok, 2012), or to a multi-talker (dialogue, conversation) context (Pickering & Garrod, 2014).

1.4 NEURAL CORRELATES OF AUDITORY FEEDBACK PROCESSING

In recent years, researchers have started to investigate the neural correlates of auditory feedback processing. The main cortical areas involved in speech feedback processing are illustrated in Figure 1.2, with arrows showing (direct or indirect) connections between them, as hypothesized by the DIVA model (Guenther, 2016). Note that other areas are involved in speech motor control (and the main theoretical models) as well, including

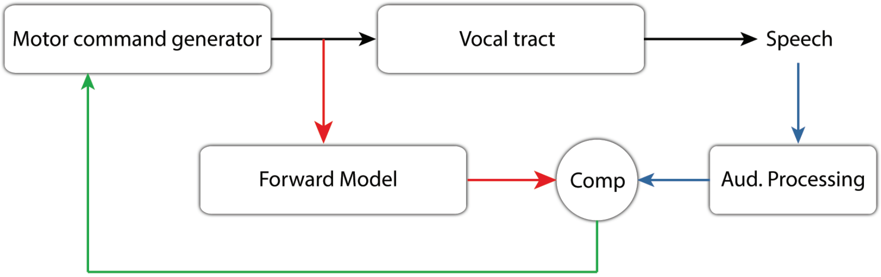

Fig. 1.1. Schematic of dominant views on feedback processing in speech motor control. Black arrows indicate a feedforward information flow from the motor system to the vocal tract, producing speech. Red arrows indicate predictive information flow, showing the efference copy being sent from the motor system to an internal forward model, leading to a sensory prediction. Blue arrows indicate basic auditory processing of the feedback signal. The observed auditory feedback is compared with the sensory prediction ("Comp"). In case of a mismatch, an error signal may be sent to the motor system (green arrow).

additional cortical areas like the right ventral premotor cortex and the supplementary motor areas, as well as subcortical areas like the basal ganglia, the thalamus and the cerebellum.

One line of research has suggested that auditory input during speech production, and auditory feedback specifically, is processed differently than auditory input when no speech is produced (Christoffels, Formisano, & Schiller, 2007; Curio, Neuloh, Numminen, Jousmaki, & Hari, 2000; Houde, Nagarajan, Sekihara, & Merzenich, 2002; Numminen, Salmelin, & Hari, 1999). These studies show reduced auditory cortex activation during speech production when compared to listening to tape recordings of the same auditory input, a phenomenon called speaking-induced suppression (SIS). The authors suggested that SIS is evidence of forward modeling in speech production: the forward model predicts the auditory consequences of articulation, and this prediction cancels out matching incoming auditory

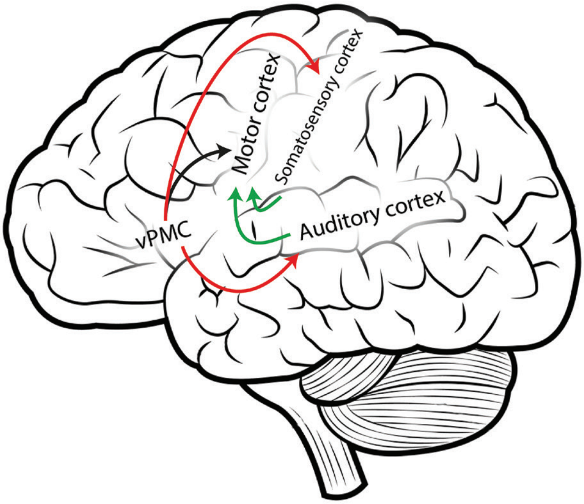

Fig. 1.2. Illustration of the main cortical areas involved in speech motor control and their connections as suggested by the DIVA model. Note that this illustration is much simplified and leaves out a number of involved areas. The ventral premotor cortex sends feedforward commands (black arrow) to the motor cortex. In addition, it sends efference copies (red arrows) to auditory and somatosensory cortices. Green arrows indicate feedback information flow. In the case of unexpected or mismatching sensory feedback, auditory and/or somatosensory cortices send information to the motor cortex, possibly mediated by other areas (not shown).

feedback, leading to reduced auditory activity.

Perturbed auditory feedback was shown to reduce or eliminate SIS, and a study using electrocorticography (Chang, Niziolek, Knight, Nagarajan, & Houde, 2013) showed that this reduction in SIS may actually be a summation of SIS and speech perturbation-response enhancement (SPRE), which do not necessarily overlap in terms of neural populations (but are summed at the scalp level). The authors suggested SIS may be related to the distinction between self- and other-generated auditory input, while SPRE may be more related to error-monitoring in speech. Findings of enhanced cortical responses to perturbed auditory feedback are line with electroencephalography (EEG) studies using an altered auditory feedback paradigm (Behroozmand, Karvelis, Liu, & Larson, 2009; Hawco, Jones, Ferretti, & Keough, 2009), which show increased auditory neural activity when auditory feedback was unexpectedly perturbed mid-production. These findings are in line with functional magnetic resonance imaging (fMRI) data as well (Behroozmand et al., 2015; Niziolek & Guenther, 2013; Tourville et al., 2008) and have led modelers to posit the locus of the comparison between expected and observed auditory feedback in auditory areas in the posterior superior temporal cortex (Guenther & Vladusich, 2012; Houde & Nagarajan, 2011). Overall, both DIVA and SFC hypothesize feedback comparison to take place in sensory areas, after which the errors are transformed to corrective movement commands that activate the pre-motor and motor cortices (Guenther, 2016; Kort, Cuesta, Houde, & Nagarajan, 2016, see also Figure 1.2).

1.5 METHODOLOGY

In order to investigate speech perception-production interactions, the research described in this thesis make use of a number of different methods. The main two techniques being used are introduced below. They are (1) altered auditory feedback and (2) magnetoencephalography. In addition, some chapters make use of still other techniques, which are described in detail in the relevant chapters. Chapter 2 uses a discrimination task with staircase procedure to quantify participants' auditory acuity (Gerrits & Schouten, 2004; Perkell et al., 2008). The staircase procedure allows us to hone in on a predetermined part of the participant's psychometric curve (Kaernbach, 1991; Levitt, 1971). Chapter 6 and 7 make use of the audiovisual recalibration paradigm (Bertelson, Vroomen, & De Gelder, 2003; van Linden & Vroomen, 2007). In this paradigm, participants make implicit use of visual cues (lip-reading) to adapt their auditory categories. The technique is well-suited to have participants shift their phoneme boundary in an implicit manner.

Most of the studies reported in the current thesis investigate the role of auditory feedback by investigating how speakers respond to experimentally altered auditory feedback (chapters 3, 4, 5 and 7). With this technique, speakers are equipped with a microphone and headphones (or audio tubes). Their speech is being recorded and sent to a dedicated sound card for manipulation. The resulting manipulated sound is played back to them in real-time (i.e., with minimal, non-noticeable, delay). The purpose of this technique is to trick the speaker into thinking they said something different to what they actually said. The manipulations used in the current thesis involve either pitch or formant values (F1 or F2). In other words, while speaking, speakers will hear themselves producing speech at a slightly different pitch than they actually did, or with slightly different formant values. The manipulations in this thesis were applied using Audapter, a software solution running on a dedicated sound card developed by Shanqing Cai (Cai, Boucek, Ghosh, Guenther, & Perkell, 2008; Tourville, Cai, & Guenther, 2013). In short, the technique applies pitch- and formant-tracking algorithms to process the recorded speech in real-time. The resulting digital signal can be used to resynthesize the sound in a way that allows for manipulation of acoustic parameters like pitch or formant values. In order for this technique of altered auditory feedback to work properly, it is important to make sure that the manipulated feedback is not overshadowed by the actually produced speech, for example via air- or bone-conducted sound. To this end, the auditory feedback was set to be rather loud, to dominate any air- or bone-conducted sounds, and when possible closed headphones were used for auditory stimulation.

Note that the altered auditory feedback technique is quite different from the imitation paradigm used in chapter 8, although both involve auditory stimulation during speech production. In chapter 8, we utilize an imitation task, where participants are instructed to start vocalizing the vowel /e/. As soon as they start speaking, one out of five vowels is played to them, and the participants' task is to change their vocalization to match the vowel they hear. Crucially, this differs from altered auditory feedback as described above because (1) the auditory stimulation is not an online manipulation of the speakers' output but a pre-recorded vowel, (2) the speakers were explicitly aware that the auditory stimulus could be one out of five Dutch vowels and (3) speakers were explicitly instructed to change their articulation. As online auditory feedback is but one example of perception-production interactions in speech, the speech imitation task was used to investigate another case of online speech sensorimotor interactions.

The neural correlates of auditory feedback processing are investigated in chapters 5 and 8 of the current thesis using magnetoencephalography (MEG). This technique is registers the electromagnetic brain response while participants are performing an experimental task (da Silva, 2010). When a neuron receives synaptic input from another neuron, postsynaptic currents are elicited in the neuron's dendrite, which may lead to this neuron being activated to send action potentials. The main contribution to the signal picked up by the MEG are the postsynaptic dendritic currents in pyramidal neurons. The parallel alignment of these neurons in the neocortex allows for summation of electromagnetic signals of thousands of neurons and thus for detection by the MEG sensors. However, some neural activity is not picked up by the MEG system (Hillebrand & Barnes, 2002). This is the case because (1) some neurons are not arranged in a way that allows for summation of their electromagnetic responses, (2) the electromagnetic signal drops off quickly with distance, meaning neural activity deep in the brain cannot be picked up by MEG and (3) MEG is not sensitive to activity in neurons with a radial orientation relative to the head surface. It is therefore important to realize that MEG picks up only part of the brain activity. Despite these limitations, MEG nevertheless comes with some advantages compared to fMRI or EEG data. While fMRI has very low temporal resolution, EEG has very low spatial resolution. MEG however, has a temporal resolution on the order of milliseconds, which is the same order of magnitude as EEG. While the spatial resolution of MEG is not as good as that of fMRI, it is better than that of EEG. An important advantage of MEG over related methods like EEG is its combination of very good temporal resolution with reasonable spatial resolution. As speech is a rapidly changing signal, MEG is well-suited to investigate its neural correlates.

1.6 THESIS OVERVIEW

This thesis reports several experimental studies that examine the influence of perception on speech production, and more specifically the role of auditory feedback in speech motor control. The influence of perception on speech production can be examined at multiple timescales. At a long timescale, dominant views of speech motor control hypothesize that if speech perception and speech production interact, over time this would cause individual variability in speech perception and speech production to co-vary. This hypothesis is examined in **chapter 2**. More specifically, we investigate whether individual variability in perceptual acuity is associated with variability in speech production precision. At shorter timescales, we investigate the role of auditory feedback during speech production. Often, a distinction is made between immediate auditory feedback processing, where unexpected or mismatching auditory feedback can lead to compensatory articulations ("compensation"), and feedback-based speech adaptation, where auditory feedback may drive sensorimotor learning by adapting the internal forward models. Although there have been studies looking at both types of auditory feedback processing in the last decades, relatively few studies have looked at their relationship. In **chapter 3**, we investigate this relationship more closely. Specifically, we hypothesized that the consistency of feedback perturbations may determine whether feedback is only for generating compensatory responses, or also for more long-term adaptations of the internal models.

As discussed earlier, it has been reported that sometimes, instead of compensating for altered auditory feedback, speakers alter their speech output by following the direction of the feedback manipulation. As it is currently unclear what the cause of this following behavior is, **chapter 4** explores whether the speech system's state at the perturbation onset may determine whether speakers oppose or follow a feedback pitch manipulation.

A secondary aim of the work reported in this thesis was to examine the neural correlates of auditory processing during speech production. Earlier research has suggested that predicted and observed auditory feedback are compared in auditory cortices, which may interact with motor cortices to implement subsequent behavioral responses. **Chapter 5** uses a pitch perturbation paradigm while measuring MEG to investigate the neural correlates of feedback processing. This chapter investigates both evoked as well as induced neural activity changes that reflect processing unexpected auditory feedback. The study tests whether feedback perturbation indeed activates error processing mechanisms in auditory and motor cortices.

The next chapters of this thesis investigate other perception-production relationships in speech. As the results in chapter 2 suggested that perception-production associations in terms of individual variability are rather weak, **chapter 7** investigates a recently proposed hypothesis suggesting that perceptual learning influences speech motor learning (rather than speech production per se). Specifically, we investigated whether audiovisual recalibration of vowel categories influenced subsequent motor learning in the form of a feedback-based speech adaptation paradigm. As a lead-in to this chapter, **chapter 6** examines whether participants actually use audiovisual information to recalibrate vowel categories.

Another case of speech perception-production interaction at a short time scale (besides auditory feedback) is speech imitation. Since the main question of the current thesis is about perception-production interactions in speech, it is important to look beyond auditory feedback during speech production.

Speech imitation is a useful case to examine in this respect, as it involves an explicit perceptual stimulus (the imitation target) that leads to behavioral responses. **Chapter 8**, instead of focusing on auditory feedback, examines the neural correlates of processing externally generated auditory signals during speech production in a speech imitation task. Neural evidence on the perception-production link in speech has suggested, relative to passive listening, reduced (or, at least, different) auditory processing during speech production. Chapter 8 tests this hypothesis more closely by extending it to an imitation task, where the auditory input is the imitation target rather than actual auditory feedback.

Finally, **chapter 9** summarizes the main findings reported in this thesis and will attempt to draw conclusions for current views on the role of auditory feedback, and speech perception more generally, during speech production.

REFERENCES

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language, 89*, 23–36. doi:10.1016/j.jml.2015.10.008

Bauer, J. J., Mittal, J., Larson, C. R., & Hain, T. C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: an automatic mechanism for stabilizing voice amplitude. *The Journal of the Acoustical Society of America, 119*, 2363–2371. doi:10.1121/1.2173513

Beddor, P. S. (2015). The relation between language users' perception and production repertoires. In *Proceedings of the 18th Congress of Phonetic Sciences*. Glasgow, UK.

Behroozmand, R., Karvelis, L., Liu, H., & Larson, C. R. (2009). Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. *Clinical Neurophysiology, 120*(7), 1303–1312. doi:http://dx.doi.org/10.1016/j.clinph.2009.04.022

Behroozmand, R., Korzyukov, O., Sattler, L., & Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: evidence for different mechanisms of voice pitch control. *The Journal of the Acoustical Society of America, 132*(4), 2468–77. doi:10.1121/1.4746984

Behroozmand, R., Shebek, R., Hansen, D. R., Oya, H., Robin, D. A., Howard, M. A., & Greenlee, J. D. W. (2015). Sensory-motor networks involved in speech production and motor control: an fMRI study. *NeuroImage, 109*, 418–28. doi:10.1016/j.neuroimage.2015.01.040

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science, 14*(6), 592–7. doi:10.1046/J.0956-7976.2003.PSCI_1470.X

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*(6), 3153–3161. doi:10.1121/1.423073

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In

Proceedings of the 8th Intl. Seminar on Speech Production (pp. 65–68). Strasbourg, France.

Casserly, E. D. (2011). Speaker compensation for local perturbation of fricative acoustic feedback. *The Journal of the Acoustical Society of America, 129*, 2181–2190. doi:10.1121/1.3552883

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences of the United States of America, 110*(7), 2653–2658. doi:DOI 10.1073/pnas.1216827110/-/DCSupplemental

Christoffels, I. K., Formisano, E., & Schiller, N. O. (2007). Neural correlates of verbal feedback processing: An fMRl study employing overt speech. *Human Brain Mapping, 28*(9), 868–879. doi:Doi 10.1002/Hbm.20315

Curio, G., Neuloh, G., Numminen, J., Jousmaki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Human Brain Mapping, 9*(4), 183–191. doi:Doi 10.1002/(Sici)1097-0193(200004)9:4<183::Aid-Hbm1>3.0.Co;2-Z

da Silva, F. H. L. (2010). Electrophysiological Basis of MEG Signals. In P. C. Hansen, M. L. Kringelbach, & R. Salmelin (Eds.), *MEG: An Introduction to Methods* (pp. 1–23). New York: Oxford University Press. doi:10.1093/acprof:oso/9780195307238.003.0001

Eliades, S. J., & Wang, X. Q. (2008). *Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature, 453*(7198), 1102–U8. doi:Doi 10.1038/Nature06910

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America, 70*(1), 45. doi:10.1121/1.386580

Escudero, P. (2007). Second-language phonology: the role of perception. In M. C. Pennington (Ed.), *Phonology in Context* (p. 109). doi:10.1017/S0952675707001327

Fairbanks, G., & Guttman, N. (1958). Effects of Delayed Auditory-Feedback Upon Articulation. *Journal of Speech and Hearing Research, 1*(1), 12–22. Retrieved from <Go to ISI>://WOS:A1958CKC3600002

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*(3), 363–376. doi:Doi 10.3758/Bf03194885

Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Menard, L., … Perkell, J. S. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *Journal of the Acoustical Society of America, 128*(5), 3079–3087. doi:Doi 10.1121/1.3493430

Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA: The MIT Press.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301. doi:10.1016/j.bandl.2005.06.001

Guenther, F. H., & Vladusich, T. (2012). A Neural Theory of Speech Acquisition and Production. *Journal of Neurolinguistics, 25*(5), 408–422. doi:10.1016/j.jneuroling.2009.08.006

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research, 130*(2), 133–141. doi:10.1007/s002219900237

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America, 123*(5), 2825–2835. doi:10.1121/1.2897042

Hartsuiker, R. J., & Kolk, H. H. J. (2001). Error Monitoring in Speech Production: A Computational Test of the Perceptual Loop Theory. *Cognitive Psychology, 42*(2), 113–157. doi:10.1006/cogp.2000.0744

Hawco, C. S., Jones, J. A., Ferretti, T. R., & Keough, D. (2009). ERP correlates of online monitoring of auditory feedback during vocalization. *Psychophysiology, 46*(6), 1216–1225. doi:10.1111/j.1469-8986.2009.00875.x

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135–145. doi:Doi 10.1038/Nrn2158

Hillebrand, A., & Barnes, G. R. (2002). A Quantitative Assessment of the Sensitivity of Whole-Head MEG to Activity in the Adult Human Cortex. *NeuroImage, 16*(3), 638–650. doi:10.1006/nimg.2002.1102

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9469813

Houde, J. F., Kort, N. S., Niziolek, C. A., Chang, E. F., & Nagarajan, S. S. (2013). Neural evidence for state feedback control of speaking. In *Proceedings of Meetings on Acoustics* (Vol. 19, pp. 060178–060178). Acoustical Society of America. doi:10.1121/1.4799495

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*(28). doi:10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience, 14*(8), 1125–1138. doi:10.1162/089892902760807140

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America, 108*(3), 1246. doi:10.1121/1.1288414

Kaernbach, C. (1991). Simple Adaptive Testing with the Weighted up-down Method. *Perception & Psychophysics, 49*(3), 227–229. doi:Doi 10.3758/Bf03214307

Kleber, F., Harrington, J., & Reubold, U. (2012). The Relationship between the Perception and Production of Coarticulation during a Sound Change in Progress. *Language and Speech, 55*(3), 383–405. doi:10.1177/0023830911422194

Kort, N. S., Cuesta, P., Houde, J. F., & Nagarajan, S. S. (2016). Bihemispheric network dynamics coordinating vocal feedback control. *Human Brain Mapping*. doi:10.1002/hbm.23114

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*(1), 54–81. doi:10.1016/j.cognition.2007.07.013

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1493), 979–1000. doi:10.1098/rstb.2007.2154

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014). Plasticity in the Human Speech Motor System Drives Changes in Speech Perception. *Journal of Neuroscience, 34*(31), 10339–10346. doi:10.1523/JNEUROSCI.0108-14.2014

Lane, H., & Tranel, B. (1971). The Lombard Sign and the Role of Hearing in Speech. *Journal of Speech Language and Hearing Research, 14*(4), 677–709. doi:10.1044/jshr.1404.677

Lee, B. S. (1950). Effects of Delayed Speech Feedback. *The Journal of the Acoustical Society of America, 22*(6), 824–826. doi:10.1121/1.1906696

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(01), 1–75. doi:10.1017/S0140525X99001776

Levitt, H. (1971). Transformed up-down Methods in Psychoacoustics. *Journal of the Acoustical Society of America, 49*(2), 467–&. doi:Doi 10.1121/1.1912375

Liberman, A. M., & Mattingly, I. G. (1985). The Motor Theory of Speech-Perception Revised. *Cognition, 21*(1), 1–36. doi:Doi 10.1016/0010-0277(85)90021-6

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say. *Psychological Science, 25*(6), 1198–1205. doi:10.1177/0956797614529797

Liu, H., & Larson, C. R. (2007). Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex. *The Journal of the Acoustical Society of America, 122*(6), 3671–7. doi:10.1121/1.2800254

Möttönen, R., & Watkins, K. E. (2011). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology, 26*(9), 1103–1118. doi:10.1080/02687038.2011.619515

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *Journal of the Acoustical Society of America, 113*(5), 2850–2860. doi:Doi 10.1121/1.1567280

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience, 33*(29), 12090–12098. doi:Doi 10.1523/Jneurosci.1008-13.2013

Numminen, J., Salmelin, R., & Hari, R. (1999). Subject's own speech reduces reactivity of the human auditory cortex. *Neuroscience Letters, 265*(2), 119–122. doi:Doi 10.1016/S0304-3940(99)00218-9

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*(4), 2382–2393. doi:10.1121/1.2178720

Pardo, J. S. (2012). Reflections on phonetic convergence: Speech perception does not mirror speech production. *Linguistics and Language Compass, 6*(12), 753–767. doi:10.1002/lnc3.367

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-Language Perspective on Speech Information Rate. *Language, 87*(3), 539–558. doi:10.1353/lan.2011.0057

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contracts in related to their discrimination of the contrasts. *Journal of the Acoustical Society of America, 116*(4), 2338–2344. doi:Doi 10.1121/1.1787524

Perkell, J. S., Lane, H., Ghosh, S. S., Matthies, M. L., Tiede, M., Guenther, F. H., & Ménard, L. (2008). Mechanisms of Vowel Production: Auditory Goals and Speaker Acuity. In *8th International Seminar on Speech Production* (pp. 29–32). Strasbourg, France.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., … Guenther, F. H. (2004). The distinctness of speakers' (s) - (integral) contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech Language and Hearing Research, 47*(6), 1259–1269. doi:Doi 10.1044/1092-4388(2004/095)

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105–110. doi:http://dx.doi.org/10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience*, 8. doi:Artn 132 Doi 10.3389/Fnhum.2014.00132

Purcell, D. W., & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America, 120*(2), 966. doi:10.1121/1.2217714

Purcell, D. W., & Munhall, K. G. (2006b). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America, 119*(4), 2288. doi:10.1121/1.2173514

Schuerman, W. L., Nagarajan, S., McQueen, J. M., & Houde, J. (2017). Sensorimotor adaptation affects perceptual compensation for coarticulation. *The Journal of the Acoustical Society of America, 141*(4), 2693–2704. doi:10.1121/1.4979791

Schütz-Bosbach, S., Mancini, B., Aglioti, S. M., & Haggard, P. (2006). Self and Other in the Human Motor System. *Current Biology, 16*(18), 1830–1834. doi:10.1016/j.cub.2006.07.048

Shadmehr, R., & Mussa-Ivaldi, F. a. (1994). Adaptive representation of dynamics during learning of a motor task. *The Journal of Neuroscience, 14*(5), 3208–3224. doi:8182467

Shadmehr, R., & Mussa-Ivaldi, S. (2012). *Biological Learning and Control*. The MIT Press. doi:10.7551/

mitpress/9780262016964.001.0001

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *J Acoust Soc Am, 125*(2), 1103–1113. doi:10.1121/1.3058638

Tourville, J. A., Cai, S., & Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech (pp. 060180–060180). doi:10.1121/1.4800684

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981. doi:10.1080/01690960903498424

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage, 39*(3), 1429–1443. doi:http://dx.doi.org/10.1016/j.neuroimage.2007.09.054

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1483–1494. doi:10.1037/0096-1523.33.6.1483

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America, 122*(4), 2306–2319. doi:Doi 10.1121/1.2773966

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nat Neurosci, 3*(Suppl), 1212–1217.

Wolpert, D., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), 1880–1882.

2

INDIVIDUAL VARIABILITY IN SPEECH PERCEPTION AND PRODUCTION

ABSTRACT

An important part of understanding speech motor control consists of capturing the interaction between speech production and speech perception. This study tests a prediction of theoretical frameworks that have tried to account for these interactions: if speech production targets are specified in auditory terms, individuals with better auditory acuity should have more precise speech targets, evidenced by decreased within-phoneme variability and increased between-phoneme distance. A study was carried out consisting of perception and production tasks in counterbalanced order. Auditory acuity was assessed using an adaptive speech discrimination task, while production variability was determined using a pseudo-word reading task. Analyses of the production data were carried out to quantify average within-phoneme variability as well as average between-phoneme contrasts. Results show that individuals not only vary in their production and perceptual abilities, but that better discriminators have more distinctive vowel production targets – that is, targets with less within-phoneme variability and greater between-phoneme distances – confirming the initial hypothesis. This association between speech production and perception did not depend on local phoneme density in vowel space. This study suggests that better auditory acuity leads to more precise speech production targets, which may be a consequence of auditory feedback affecting speech production over time.

This chapter was based on:

Franken, M. K., Acheson, D.J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *Journal of the Acoustical Society of America*, 142(4): 2007-2018.

2.1 INTRODUCTION

How do speech perception and speech production interact? Several lines of research have shown that speech production and speech perception are not independent processes, but interact in complicated ways. Investigations of these perception-production interactions can largely be placed in two categories. The first type focuses on short-term effects of perception on production. For example, when a speaker's auditory feedback is manipulated or distorted, his or her speech production is affected (Elman, 1981; Fairbanks & Guttman, 1958; Houde & Jordan, 1998; Purcell & Munhall, 2006). For example, when auditory feedback is delayed by just 200ms, speakers make more speech errors (Fairbanks & Guttman, 1958), and, when the pitch of individuals' speech is artificially shifted up in auditory feedback, speakers compensate by shifting their pitch downward (Burnett, Freedland, Larson, & Hain, 1998). Although these studies have shown that auditory feedback is not strictly necessary for regular speech production (Lane & Webster, 1991), they also demonstrate that the perception and production systems interact in real-time.

The second line of research into the perception-production link focuses on longer-term interactions between speech production and perception, usually by studying correlations between the two. Here, the guiding hypothesis is that if production and perception interact on a daily basis, this will lead to co-variation across individuals. For example, Newman (2003) investigated correlations between acoustic measures of listeners' perceptual prototypes for a given speech category and of their average production of members of that category. People whose perceptual prototype of stop consonants had a longer voice onset time (VOT) also tended to produce these consonants with longer VOT.

Another example of research into longer-term interactions concerns studies that have shown a correlation between auditory acuity and vowel production (Perkell et al., 2004, 2008). In these studies, participants carried out two tasks, (1) a discrimination task on a vowel continuum and (2) an overt reading task. The results showed that participants who were better at the discrimination task produced vowels more consistently (less within-phoneme variability) but spaced them further apart in vowel space (larger between-phoneme acoustic distance). The authors interpret their findings as follows: better auditory acuity is reflective of more precise speech targets (e.g., smaller target regions in acoustic space), which in turn leads to more consistent speech production, as a smaller target region would result in more rejections of non-prototypical productions as 'speech errors'. A related study is reported by Villacorta et al. (2007), who showed that people with higher auditory acuity compensate more strongly in response to altered auditory feedback.

The interplay between speech production and speech perception has also been corroborated in neurobiological studies. Several studies have shown that auditory input is processed differently during speech production compared to passive listening (Christoffels, van de Ven, Waldorp, Formisano, & Schiller, 2011; Franken, Hagoort, & Acheson, 2015; Heinks-Maldonado, Nagarajan, & Houde, 2006; Houde, Nagarajan, Sekihara, & Merzenich, 2002). Both behaviorally and neurobiologically, it is well established that unexpected auditory feedback leads to subsequent changes in speech production (Behroozmand, Ibrahim, Korzyukov, Robin, & Larson, 2015; Behroozmand, Karvelis, Liu, & Larson, 2009; Parkinson, Korzyukov, Larson, Litvak, & Robin, 2013). Note, however, that the amount and significance of individual variability in these interactions is not well understood. Along with previous studies, the current study will offer an example of how studying individual variability can illuminate the interplay between perception and production in speech motor control.

Several current theories of speech motor control hypothesize that speech perception contributes to speech production through an auditory feedback mechanism that informs speech motor control (Hickok, Houde, & Rong, 2011; Houde & Nagarajan, 2011; Tourville & Guenther, 2011). Further highlighting the important link between perception and production, these models posit that speech production goals are ultimately perceptual targets. In other words, the goal of the speech production process is to produce a particular sound sequence. It is assumed that these sound representations are first acquired via speech perception, making it conceivable that speech production targets will co-vary with individual variability in speech perception. In addition, the speech production process might be tuned over time by perception: auditory feedback processing may reject the produced speech sound as a deviation from the prototypical representation, leading to compensatory responses. In such a system, individuals with higher perceptual acuity may be more sensitive to speech production that deviates from expected targets. Such deviations might be detected as speech errors, which over time would drive the production system to be more precise (i.e., less variable).

Although the above models make clear predictions about within-phoneme co-variation, it remains unclear whether the production-perception co-variation would also vary across phonemes. It is well established that vowel space is perceptually warped by the presence of phonemes (Kuhl, 1991; Kuhl et al., 2008). Therefore, it is conceivable that associations between speech perception and production may vary both locally, for example depending on the local phoneme density, as well as cross-linguistically, depending on the language's phoneme inventory. An example of local and cross-linguistic differences in phoneme inventories is shown in Figure 2.1, which depicts the vowel inventories of Dutch and English, two closely related languages. Although the two languages have a similar number of vowels, it can be seen in Figure 2.1 that in Dutch 'front' vowels (those at the higher end of the F2 scale) exist in a higher density space than the 'back' vowels, whereas this is not the case in English. This is corroborated by analyses of Dutch interphonemic distances in Figure 2.1 which showed that, for example, Dutch /ɑ/ lies in a less dense space compared to Dutch /ɛ/, both globally (overall average distance to other phonemes, /ɑ/: 732ΔHz, /ɛ/: 590ΔHz) and locally (average distance to 3 closest phonemes, /ɑ/: 424ΔHz, /ɛ/: 244ΔHz; distance to closest phoneme, /ɑ/: 330ΔHz, /ɛ/: 199ΔHz). When other phonemes are nearby, it would pay off to have very precise articulatory targets, so that the produced vowel is not confused with the neighboring phonemes. Previous research suggests that neighboring phonemes indeed have an effect on phonemes' target regions, as auditory feedback control is modulated by the presence of nearby phoneme categories (Niziolek & Guenther, 2013). It has not been shown, however, whether phoneme density also affects longer-term interactions between the perception and production systems. For example, higher phoneme density might drive the system to develop stronger perception-production links than those in lower-density regions of the acoustic space. So in denser parts of vowel space, people might be more sensitive to deviations, which would lead them to develop smaller targets. Therefore, we tested the hypothesis that the relationship of speech perception with speech production variability is affected by local phoneme density.

In the present study we address whether the longer-term production-perception interactions discussed above result in associations between perception and production behavior. More specifically, we determine whether auditory acuity, as measured by a speech discrimination task, would be associated with individual variability in vowel productions. This was done by having participants carry out a speech discrimination task and a speech production task, and investigating possible correlations of individual variability across tasks, using a similar paradigm to Perkell et al. (2008). In addition, we investigated whether these perception-production associations depend on local vowel density by comparing a pair of front vowels with a pair of back vowels (the bold labels in Figure 2.1). In terms of speech discrimination, like Perkell et al. (2008) we used a 4-interval.

2-alternative forced choice task, which has been shown to capture lower-level auditory discrimination, with relatively little influence from phonemic categories (Gerrits & Schouten, 2004). However, in the present study we measured auditory acuity using a discrimination score, a measure that takes into account both participants' overall discrimination ability as well as the consistency of their performance, whereas previous methods only captured participants' average performance. The speech production task in the present study was a non-word reading task. In order to characterize production variability, measures were used that take into account distributional properties of vowel space as well as measures that capture psychophysical properties of speech perception. For example, in addition to characterizing vowel production in terms of F1 and F2 measurements (as done in most research in this area), we also characterized vowel production in terms of so-called mel-frequency cepstral coefficients (MFCC). These coefficients represent spectral properties of speech and are widely used in the field of automatic speech recognition. In contrast to F1/F2 values, they provide a broader representation of the spectral shape of speech sounds and

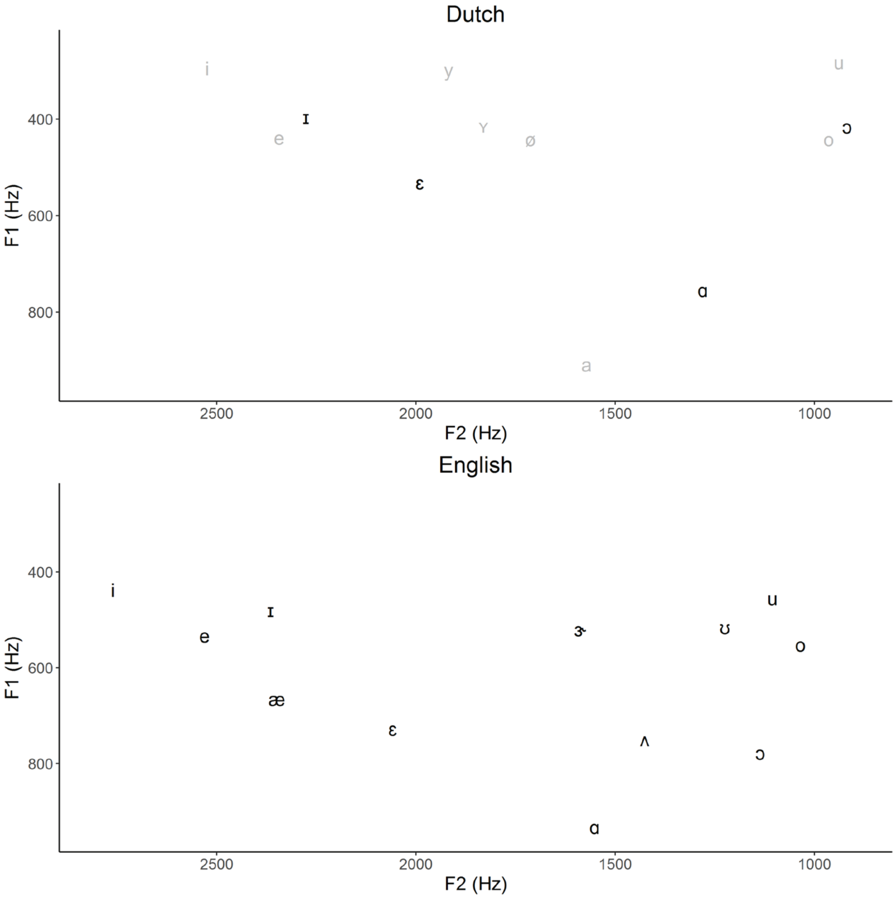

Fig. 2.1. The vowel spaces of Dutch and English, exemplified by vowels plotted as a function of average first (F1) and second (F2) formant values. Dutch data shown here are the acoustic values of vowels spoken by females speaking Northern Standard Dutch, reported in Adank et al. (2004), excluding the three Dutch diphthongs /ɛɪ/, /ɑu/ and /œy/. Bold labels indicate the vowels we used in our study (/ɪ/, /ɛ/, /ɑ/, /ɔ/). English data are acoustic measurements of female speakers of American English, taken from Hillenbrand et al. (1995).

are designed to better reflect the psychophysics of human vowel perception. Compared to the acoustic measures used in Perkell et al. (2008), which rely on Euclidian distances in 2-D vowel space, we believe these measures are better able to capture human speech perception.

If the models of speech production mentioned earlier are correct, long-term interactions between perception and production should lead to co-variability across individuals and thus we expect our perception measures to correlate with speech production variability. More specifically, these models predict that individuals with better auditory acuity would have more precise vowel targets and therefore show less production variability. In addition, we investigate whether these predicted associations between perception and production vary as a function of local phoneme density. Therefore we will compare higher density Dutch front vowels /ɪ/ and /ɛ/ (from a denser part of vowel space) with Dutch back vowels /ɑ/ and /ɔ/ (from a sparser part of vowel space).

2.2 METHODS

2.2.1 Subjects

Forty healthy volunteers (age: M = 20, SD = 2.2; 24 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local ethics committee (the Social Sciences Ethical Committee of Radboud University). All participants had normal hearing, were native speakers of Dutch and had no history of speech and/or language pathology. Three reported being raised multilingually, the others were raised monolingually in Dutch (though seven reported speaking a local dialect). All participants also reported how many languages they learned (at school or elsewhere) aside from Dutch. As is common in the Netherlands, most of them reported having learned three languages besides Dutch (M=3, SD = 0.92, range = 2-5).

2.2.2 Stimuli

For the discrimination task, two speech continua were created based respectively on recordings of the pseudowords skef and skaf, spoken by a male native Dutch speaker. From each of these recordings, the two continua (/skɛf/-/skɪf/ and /skɑf/-/skɔf/) were made by manipulating F1 and F2 values. First, the vowels were excised from each recording. Using Burg's linear predictive coding (LPC) framework, a filter model was obtained by estimating five formants between 0 and 5000Hz. A source model was obtained using eight prediction coefficients. A number of filter models were created by changing F1/F2 values in a stepwise manner, and the endpoints of the continua were based on the average F1 and F2 values for a male Dutch speaker (Adank, van Hout, & Smits, 2004), as these came close to the values of the original recording. For the skaf-skof continuum, 1001 steps were used (as in Perkell et al., 2008), each one having a change of -0.176Hz in F1 and -0.351Hz in F2. For the skef-skif continuum, 543 steps were created, so the Euclidian distance in F1*F2 space between successive steps was similar to the first continuum (F1 change was -0.210Hz, F2 change was 0.332Hz). This allowed us to compare results on both continua. These filter models were combined with the source model. The results were lowpass-filtered at 2000Hz and combined with the band-pass filtered original signal (2000Hz-6000Hz). This way, it was ensured that above 2000Hz, the signal was exactly the same as the original. All vowels were manipulated so their average intensity matched that of the original sounds. Finally, the vowels were embedded in the sk_f context, which was exactly the same for all stimuli in a continuum (the consonantal frame was taken from the original pseudoword recording).

For the production task, pseudowords were created using a C1V1C1C1V1C2 structure, where C1 could be either one of /k/, /p/ or /t/, V1 either one of /ɛ/, /ɪ/, /ɑ/ or /ɔ/, and C2 either one of /p/, /t/, /k/, /f/, /s/ or /x/. This particular structure was used because monosyllabic structures led to too many existing words (rather than pseudowords), and the various consonants used were all voiceless obstruents, making it easier to later determine vowel onsets and offsets in the recordings. Using all possible combinations of these vowels and consonants resulted in 72 unique pseudowords (e.g., kekkef, poppos).

2.2.3 Procedure

The experiment consisted of two tasks, which were administered in counter-balanced order within a single session with a short break in between.

The discrimination task consisted of a four-interval two-alternative forced choice task (Gerrits & Schouten, 2004) with a staircase technique based on the weighted up-down procedure (Kaernbach, 1991; Levitt, 1971). On every trial subjects heard four auditory stimuli: three standard stimuli and one deviant stimulus. The standard stimuli were always one extreme of the continuum (i.e., three times the same stimulus, skef for the skef-skif continuum, skaf for the skaf-skof continuum), while the deviant stimulus varied on a trial-by-trial basis. The deviant stimulus occurred in position two or three, and the participant was instructed to push the left button when he or she thought the deviant was the second stimulus, and to push the right button when he or she thought it was the third stimulus. If the participant responded correctly, the difference between the standard and the deviant in the next trial was decreased, otherwise it was increased. Participants did not receive feedback on their performance.

The discrimination task was divided into four blocks, which alternated between continua. Every block started with a fairly large interval (250 continuum steps or Euclidian distance in F1*F2 space of around 98.2ΔHz between standard and deviant stimulus). 'Reversal' trials were trials where subjects gave a correct response after a previous incorrect trial, or vice versa. The block ended after a total of 20 reversal trials. The amount of change in the interval size from trial to trial was initially large (a decrease of 25 steps after a correct trial, an increase of 75 after an incorrect trial), and became smaller after the second reversal trial of a block (a decrease of 10 after a correct trial, an increase of 30 after an incorrect trial). Because the increase in interval size after an incorrect trial was always three times the decrease of the interval size after a correct trial, the interval size should theoretically converge to a threshold interval size where people would give a correct answer on 75% of the cases (Kaernbach, 1991).

The production task was a simple pseudoword reading task. Subjects were instructed to read aloud the pseudowords that appeared on the screen, while trying to maintain a constant, normal volume and making sure stress was placed on the second syllable (which was printed in capitals). Subjects were positioned about 30 cm from the microphone and asked to try to keep this distance throughout. The task consisted of four blocks, each of which presented all 72 pseudowords in randomized order. Every pseudoword was thus repeated four times.

2.2.4 Hardware

All recordings were made in a soundproof booth and digitized at 44.1 kHz on one channel using a Sennheiser ME64 microphone, which was set up in the booth and connected through an Alesis Multimix 6 FX audio mixer to a Windows computer outside the booth. Auditory stimuli were delivered through the same audio mixer which was connected to Sennheiser HD280-13 headphones. Stimuli presentation and sound recording times were controlled by the same Windows computer running Neurobehavioral Systems Presentation.

2.2.5 Analysis

2.2.5.1 Perception

For the results from the discrimination task, we calculated a threshold value per block by averaging the interval sizes for the last 16 reversal trials. Subsequently, we took the minimal threshold per continuum for each subject. As another measure of discrimination performance, we quantified the consistency between blocks of the same continuum in the following way: we created a linear mixed effects model, with Block and Continuum as fixed effects, Subject as a random effect (with random slopes for Block and Continuum) and the calculated thresholds as dependent variables. The absolute values of the random slopes for Block were taken as a measure of between-block inconsistency. Finally, we also calculated a "discrimination score" by multiplying the between-block inconsistency measure by the minimal threshold value. So the discrimination score could be high either because the participant was not very consistent between blocks, or had a high minimal threshold. In other words, a higher discrimination score corresponds to worse performance on the discrimination task.

We also carried out a correlation analysis between the minimal threshold and between-block inconsistency measures, in order to characterize the relationship between these two measures.

2.2.5.2 Production

For all recordings, the beginning and ending of the vowel in the second syllable, which always carried stress, was manually determined. Then the duration and formant values were extracted. Formant values were calculated by averaging over a 40ms time window at the center of the vowel. Five formants were estimated between 0 and either 5kHz (males) or 5.5kHz (females) using an iterative Burg algorithm in Praat (Boersma & Weenink, 2013). Even though in the present study we were only interested in F1 and F2, estimating five formants tends to give a more reliable result (Boersma & Weenink, 2013). For all further analyses, formant values were converted from Hertz to the Bark scale (which is defined so that the critical bands of human hearing all have the width of one Bark; Zwicker, 1961).

In order to capture subjects' production variability, two different measures were taken. The first was vowel dispersion, or the area of the ellipse described by one standard deviation in both F1 and F2 for that phoneme. This was calculated using the formula of the area of the ellipse:

$$\textit{Vowel Dispersion} = \pi \cdot x \cdot y$$

Here, x and y correspond to one standard deviation in F1 and F2 respectively. This corresponds to what others have called "compactness score" (Kartushina & Frauenfelder, 2013, 2014). Vowel dispersion was calculated per vowel, and the results were averaged across vowels within subjects. The second measure was average vowel spacing (AVS), which was the average Mahalanobis distance between the phoneme's centroid and all neighboring phoneme distributions. This was averaged across all possible vowel pairings (i.e., between /ɪ/ centroid and /ɛ/ distribution, /ɛ/ centroid and /ɪ/ distribution, /ɪ/ centroid and /ɔ/ distribution, etc.). A similar measure was also used by Kartushina & Frauenfelder (2013, 2014). Both dispersion and AVS were calculated in F1*F2 space.

Similar analyses were conducted using mel-frequency cepstral coefficients (MFCCs; Gold, Morgan, & Ellis, 2011). MFCC representations mimic the workings of the filter bank in the inner ear. MFCC calculations were done in Praat by first performing a filter bank analysis with 12 filters (first filter centered at 100 mel, distance between successive filters 100 mel). Subsequently, the filter values were converted to MFCCs using a Discrete Cosine Transform. Finally, dispersion was quantified as the mean Euclidian distance to the centroid in 12-dimensional space (defined by 12 MFCCs), and AVS as the average pairwise distance between vowel centroids in the 12-dimensional MFCC space.

2.2.5.3 Perception vs. production

In order to assess the association between perception and production variability, regression analyses were carried out with discrimination score (as defined in the section on the analysis of the perception data above) as the dependent variable and the production measures as well as vowel continuum as the predictors. Data points for which Cook's distance was larger than 0.1 for a particular analysis (indicating high residuals and/or high leverage) were removed from that analysis (on average 3.25% of the data points were removed).

2.3 RESULTS

2.3.1 Discrimination

For every participant, the discrimination threshold was calculated for every block in both continua. The results for a representative participant are shown in Figure 2.2. Although the average threshold across subjects for both continua was lower in the second block (/ɛ/-/ɪ/ block 1: M = 83.1Δbark (SD = 53.1), block 2: M = 72.2Δbark (42.4); /ɑ/-/ɔ/ block 1: M = 128.5Δbark (62.9), block 2: M = 105.8Δbark (50.3)), there were also subjects who showed an increased threshold for both continua in the second block (17 subjects for /ɛ/-/ɪ/, 13 subjects for /ɑ/-/ɔ/). If we take the minimum threshold for each participant and each continuum, we see that the /ɑ/-/ɔ/ continuum was harder than the /ɛ/-/ɪ/ continuum (/ɛ/-/ɪ/: M = 60.2Δbark (31.3); /ɑ/-/ɔ/: M = 97.4Δbark (48.3)). This difference was significant (*t*(76.71) = -4.86, $p < 0.001$, t-test done on log-transformed threshold values).

With respect to within-subject variability, there were positive correlations between participants' discrimination threshold in the first block and their threshold in the second block for both continua (for /ɛ/-/ɪ/: *r*(38) = 0.63, $p < 0.001$ and for /ɑ/-/ɔ/: *r*(38) = 0.59, $p < 0.001$). Although a positive correlation was expected (given that participants performed the same task on the same stimuli), it explained only about 40% and 36% of the variability, respectively, indicating that participants did not perform consistently in either block. To quantify this variability, we performed a linear mixed effects model analysis on the participants' thresholds with Continuum (/ɛ/-/ɪ/ vs. /ɑ/-/ɔ/) and Block (block 1 or block 2) as fixed effects and random slopes within subjects. The results are shown in Table 2.1.

As a measure of participants' inconsistency in their performance, we took the absolute value of the random effects for the Block predictor. For both continua, this inconsistency value correlated weakly with the participants' minimal thresholds (/ɛ/-/ɪ/: *r*(37) = 0.24; /ɑ/-/ɔ/: *r*(37) = 0.23; see Figure 2.3). In other words, participants with a higher minimal threshold (worse discrimination performance) also performed less consistently in the discrimination task.

In order to quantify both performance inconsistency and discrimination threshold, we used the discrimination score (inconsistency*threshold) in subsequent analyses.

2.3.2 Production

To quantify speech production variability, we used measures in F1*F2 space, as the majority of research in acoustic phonetics characterizes vowel acoustically in terms of formant values. We also used measures in mel-frequency cepstral coefficient (MFCC) space. MFCC values are designed to capture vowel acoustics in a way that is closer to human perception, as these coefficients are based on filter banks similar to known variation of the ear's critical bandwidths (Davis & Mermelstein, 1980). In both F1*F2 and MFCC domains we had a measure of within-phoneme variability and a measure of between-phoneme distance.

In F1-F2 space, the within-phoneme variability measure (ellipse area, see Figure 2.4 for an example) had cross-participant means of 0.27Δbark (0.17) for /ɑ/, 0.25Δbark (0.10) for /ɛ/, 0.17Δbark (0.09) for /ɪ/ and 0.30Δbark (0.21) for /ɔ/ (standard deviations between brackets). For the between-phoneme distance measure (Average Vowel Spacing (AVS) or mean squared Mahalanobis distances), we find a mean of 160.9Δbark^2 (70.2). For further correlation analyses (see below), variability due to gender was removed from this measure, as this also affects AVS values. This was done by generating a linear model with AVS as the independent measure and a single predictor that coded for gender. The linear model showed that male speakers had smaller AVS values (i.e., a smaller vowel space) than female speakers ($F(1, 38) = 12.98$, $p < 0.001$), as is well known from the literature (Simpson, 2001, 2009). The residuals of this linear model, reflecting variability in AVS that cannot be attributed to gender differences, were used as input in the correlation analyses.

Table 2.1. Linear Mixed Effects Model results, looking at discrimination thresholds in terms of Block and Continuum, with random slopes for both within Subjects.

	estimates	t-values		estimates
Fixed effects			*Random effects*	
Intercept	86.002 (7.998)	10.752	Intercept (Subjects)	2037.4
Block	-16.736 (6.000)	-2.789	Block (Subjects)	744.6
Continuum	39.508 (8.458)	4.671	Continuum (Subjects)	2166.3
Residual	695.4			

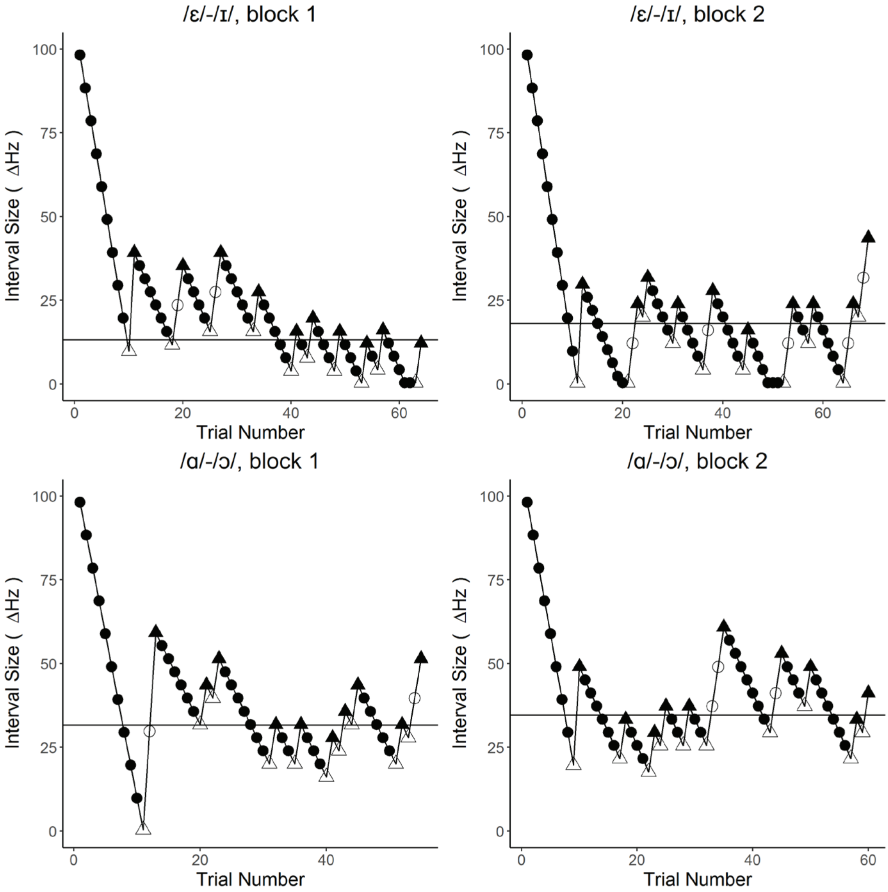

Fig. 2.2. Discrimination results for a representative participant. Every panel of the plots shows the interval size as a function of trial number for a particular experimental block. The top row shows the two blocks for the /ɛ/-/ɪ/ continuum, the bottom row those for the /ɑ/-/ɔ/ continuum. The left column shows the first block for each continuum and the right column the second block. The solid symbols indicate trials that were answered correctly, empty symbols indicate trials where the response was incorrect. Triangles indicate reversal trials. The horizontal line indicates the threshold calculated for that block.

2.3.3 Production-Perception Associations

Regression analyses were performed in order to compare individual variability in the discrimination and production tasks. Specifically, participants' perceptual Discrimination Scores were used as the dependent variable with Vowel Continuum (/ɛ/-/ɪ/ and /ɑ/-/ɔ/) as a predictor and two production-based predictors: Dispersion (within-phoneme variability; vowel ellipse area) and AVS (between-phoneme distance). These analyses were run twice, once with the F1-F2 production measures and once with the MFCC measures.

For the production measures in F1-F2 space, we first ran a full model including the predictors Dispersion, AVS, and Vowel Continuum, as well as the latter's interaction terms with both production measures. The results of the regression analysis (after having removed four data points for which Cook's distance was over 0.1) are shown in Table 2.2.

As is shown in Table 2.2, none of the interaction terms is significant, showing that any association between the production and the perception measures is not dependent on the vowel. Next, the same variables were entered in a stepwise regression procedure. This procedure allowed us to arrive at a model in which predictors that do not significantly increase

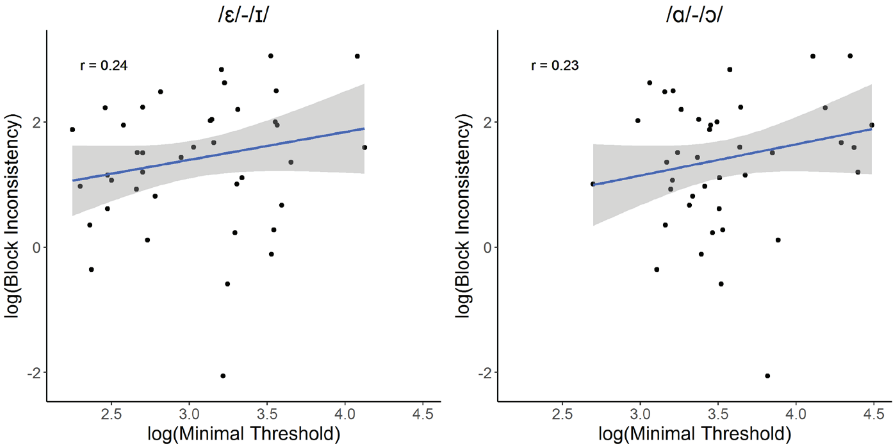

Fig. 2.3. Scatter plots of the association between participants' Minimal Threshold and their Block Inconsistency score (both log-transformed) for the /ɛ/-/ɪ/ continuum (left) and for the /ɑ/-/ɔ/ continuum (right). The superimposed line represents the best linear fit, ignoring the outlier at the low end of the Block Inconsistency measure. Shading represents the 95% confidence interval.

the model's goodness of fit were left out. The outcome of this procedure suggested the best final model included both production terms (Dispersion and AVS) as predictors, excluding Vowel Continuum as well as the interaction terms. The results of this final model are shown in Table 2.3.

The results show a significant (negative) main effect of AVS, suggesting that better performance in the discrimination task (i.e., lower discrimination score) was associated with larger between-phoneme distances in production. In other words, people who were better in speech discrimination produced vowels that were spaced further apart in vowel space. In addition, the main effect of vowel dispersion is marginally significant, indicating that people who produce vowels with less within-phoneme variability perform better at the discrimination task. However, this should be interpreted with caution given that this effect is only marginally significant in the final model, and not significant in the full model (see Table 2.2). The absence of interaction terms

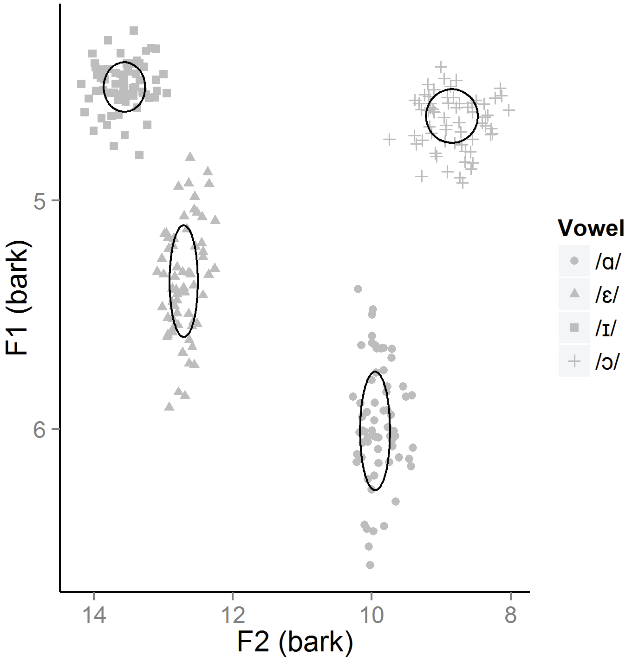

Fig. 2.4. Production data from a representative participant. The grey symbols show single trial results in terms of F1 and F2 (both in bark). Symbol shapes indicate the vowel; ellipses show the within-phoneme variability measure (area of the ellipse) for each vowel phoneme.

between any production measure and Vowel Continuum confirms that the association between production and performance in the discrimination task does not depend on specific vowels.

The pattern of results as shown in the regression analyses is in line with the pairwise correlation analyses per vowel continuum, shown in Figure 2.5. We found positive correlation coefficients for the comparison between Vowel Dispersion (i.e., within-phoneme variability) and Discrimination Score (/ɑ/-/ɔ/: *r*(36) = 0.38, *p* = 0.02, Figure 2.5 top right and /ɛ/-/ɪ/: *r*(36) = 0.25, *p* = 0.14, Figure 2.5 top left). This confirms the finding of the regression analyses that better discrimination performance (i.e., a lower Discrimination Score) was associated with less within-phoneme variability, that is, more precise vowel production. For the between-phoneme distance measure, or AVS, we found negative correlation coefficients (For /ɑ/-/ɔ/: *r*(37) = -0.33, *p* = 0.04, Figure 2.5 bottom right, and for /ɛ/-/ɪ/: *r*(37) = -0.19, *p* = 0.25, Figure 2.5 bottom left). Again, this confirms the regression results, showing that better discrimination performance was associated with larger between-phoneme distances. In other words, speakers who were better at the discrimination task produced vowels that were further apart in vowel space. Although not all significant, the scatter plots in Figure 2.5 are similar across vowel continua (and the correlations have the same direction), which is in line with the lack of interactions with Vowel Continuum in the regression analyses being significant. As above, these results are consistent with an association between speech perception and production that does not dependent on the

Table 2.2. Regression coefficients for the full model with production measures in F1*F2 space.

	Estimates	t-values	p-values
Intercept	9.94 (1.47)	6.77	<0.0001*
AVS[a]	-0.71 (0.30)	-2.39	0.019*
Dispersion	0.31 (0.67)	0.46	0.65
Vowel Continuum	1.74 (2.25)	0.77	0.44
AVS : Vowel Continuum	-0.11 (0.41)	-0.27	0.78
Dispersion : Vowel Continuum	0.62 (0.91)	0.68	0.50

[a]AVS = Average Vowel Spacing

vowel (and thus on local vowel density in vowel space).

Note that in the analyses reported here, the production measures were calculated across the entire vowel space, in contrast to the discrimination performance (which was continuum-specific). This was done to get more reliable estimates of people's phoneme dispersion and average between-phoneme spacing. Production variability can be affected by various factors, and averaging across phonemes generates in our view a better estimate of overall within-phoneme variability and between-phoneme distinctions. Consistent with the assumption that the production data for individual phonemes is more variable, there is no significant association for Vowel Dispersion (/ɑ/-/ɔ/: $r(36) = 0.20$, $p = 0.23$, and /ɛ/-/ɪ/: $r(36) = 0.20$, $p = 0.24$, or for AVS (/ɑ/-/ɔ/: $r(36) = -0.15$, $p = 0.36$, and /ɛ/-/ɪ/: $r(36) = 0.20$, $p = 0.22$) if Vowel Dispersion and AVS are computed separately for each continuum. Note that, for dispersion, this was done by averaging the dispersion values for the two endpoints of each continuum, to correspond with the discrimination measures because they necessarily involve both vowels.

Similar results were found when using the production measures in MFCC space. Table 2.4 shows the results of a full regression model, including MFCC production measures Dispersion and AVS, as well as the Vowel Continuum term and its interactions with the production measures (after having removed two data points for which Cook's distance was over 0.1).

Similar to the results above, none of the interaction terms with Vowel Continuum is significant. Next, the same variables were entered in a stepwise regression procedure. The outcome of the stepwise regression suggested that the best final model included both production terms (Dispersion and AVS) as predictors, as well as Vowel Continuum, but excluding the interaction terms.

Table 2.3. Regression coefficients for the final model with production measures in F1*F2 space.

	Estimates	t-values	p-values
Intercept	10.97 (1.06)	10.33	<0.0001*
AVS[a]	-0.73 (0.20)	-3.73	0.00038*
Dispersion	0.79 (0.40)	1.96	0.053

[a] AVS = Average Vowel Spacing

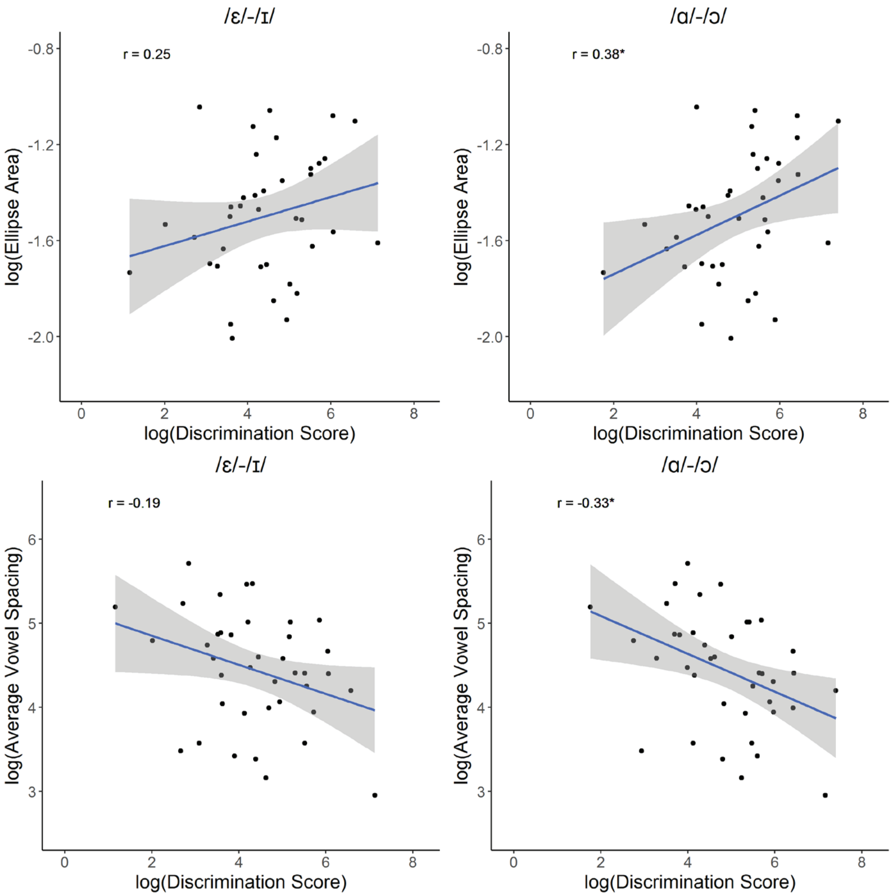

Fig. 2.5. Scatter plots of correlation analyses in F1*F2 space. Top row shows comparisons between discrimination score (x-axis) and within-phoneme variability (y-axis). Bottom row shows comparisons between discrimination score and average between-phoneme distance (y-axis). Left column shows results for the /ɛ/-/ɪ/ continuum, the right column for the /ɑ/-/ɔ/ continuum. Superimposed lines represent the best linear fit, shading represents 95% confidence interval.

The results of this final model are shown in Table 2.5.

The results of the final model show significant main effects of AVS, Dispersion and Vowel Continuum. The main effects of AVS and Dispersion confirm the effects already found in the full model (Table 2.4). These effects suggest that better auditory speech discrimination is associated with smaller within-phoneme variability (less dispersion) in production as well as larger between-phoneme distances. In other words, better discriminators produce vowels more precisely and space them further apart in vowel space. In addition, we find a significant main effect of Vowel Continuum, suggesting discrimination performance is worse in the /ɑ/-/ɔ/ continuum compared to the /ɛ/-/ɪ/ continuum. Although this effect should be taken with some caution, as this was not significant in the full model, it is in line with the significant difference between /ɛ/-/ɪ/ and /ɑ/-/ɔ/ discrimination performance found earlier.

The same pattern of results can be seen in the pairwise correlation analyses per vowel continua, shown in Figure 2.6. For the within-phoneme variability measure, we found a significant positive correlation for the /ɛ/-/ɪ/ continuum (*r*(37) = 0.35, *p* = 0.03, see Figure 2.6 top left), but not for the /ɑ/-/ɔ/ continuum (*r*(37) = 0.25, *p* = 0.12, see Figure 2.6 top right). The between-phoneme distance measures in MFCC space also showed negative correlation coefficients, which was significant for the /ɑ/-/ɔ/ continuum (*r*(37) = -0.34, *p* = 0.03, see Figure 2.6 bottom right), but not for the /ɛ/-/ɪ/ continuum (*r*(37) = -0.24, *p* = 0.15, see Figure 2.6 bottom left). These pairwise correlations are consistent with the results of the regression analyses and the results from the analyses of the F1*F2 space measures. The absence of any significant interaction between Vowel Continuum and the production measures shows that these kinds of production-perception association are not dependent on specific vowels. In line with this, the scatter plots in Figure 2.6 look similar across vowel continua, with correlations in the same direction and of similar magnitudes.

In order to be able to compare the present results with those reported in Perkell et al. (2008), analyses were also performed with the methods used in that study (see appendix). Associations were reported in that study between auditory acuity and both vowel dispersion and average vowel spacing. Using these analysis methods on the present data did not show statistically significant correlations, suggesting that the methods presented here (using amongst others metrics based on ellipse area, Mahalanobis distances, and MFCCs) were better able to capture these perception-production associations. The direction of the (non-significant) correlation trends was, however, in the same direction. That is, people with higher auditory acuity tended to show less vowel dispersion and more average vowel spacing.

2.4 DISCUSSION

In the present study we compared individual variability in a speech perception task with variability in speech production. Our reasoning was that if, as the literature suggests, speech production and speech perception interact over time, individual differences in these domains should correlate. The results showed that better discrimination performance was associated with less within-phoneme variability in production, as well as with larger average between-phoneme distances. This picture emerged both from the analyses using production measures in F1*F2 space as well as from the analyses using measures in MFCC space. In addition, none of the regression analyses showed a significant interaction with vowel continuum regardless of whether the target vowels were in the denser front part of vowel space or in the sparser back part. This suggests that the perception-production association was not dependent on specific vowels or local phoneme density,

Table 2.4. Regression coefficients for the full model with production measures in MFCC space.

	Estimates	t-values	p-values
Intercept	10.59 (14.66)	0.72	0.47
AVS[a]	-5.36 (2.47)	-2.17	0.033*
Dispersion	5.61 (1.98)	2.83	0.006*
Vowel_Continuum	12.74 (20.73)	0.62	0.54
AVS : Vowel_Continuum	-1.12 (3.50)	-0.32	0.75
Dispersion : Vowel_Continuum	-1.37 (2.80)	-0.49	0.63

[a]AVS = Average Vowel Spacing

Table 2.5. Regression coefficients for the final model with production measures in MFCC space.

	Estimates	t-values	p-values
Intercept	16.70 (10.25)	1.63	0.11
AVS[a]	-5.92 (1.73)	-3.42	0.0010*
Dispersion	4.92 (1.39)	3.55	0.00067*
Vowel_Continuum	0.53 (0.25)	2.11	0.039*

[a]AVS = Average Vowel Spacing

but instead holds across vowel space. This is also corroborated by the fact that when we used continuum-specific production measures, no significant associations were found.

These results are largely in line with previous findings by Perkell et al. (2004, 2008), although these earlier studies reported much stronger effects. It is unclear what drives the difference in effect sizes. Although we tested native speakers of Dutch whereas Perkell et al. tested native speakers of English, we would not expect the link between perception and production in

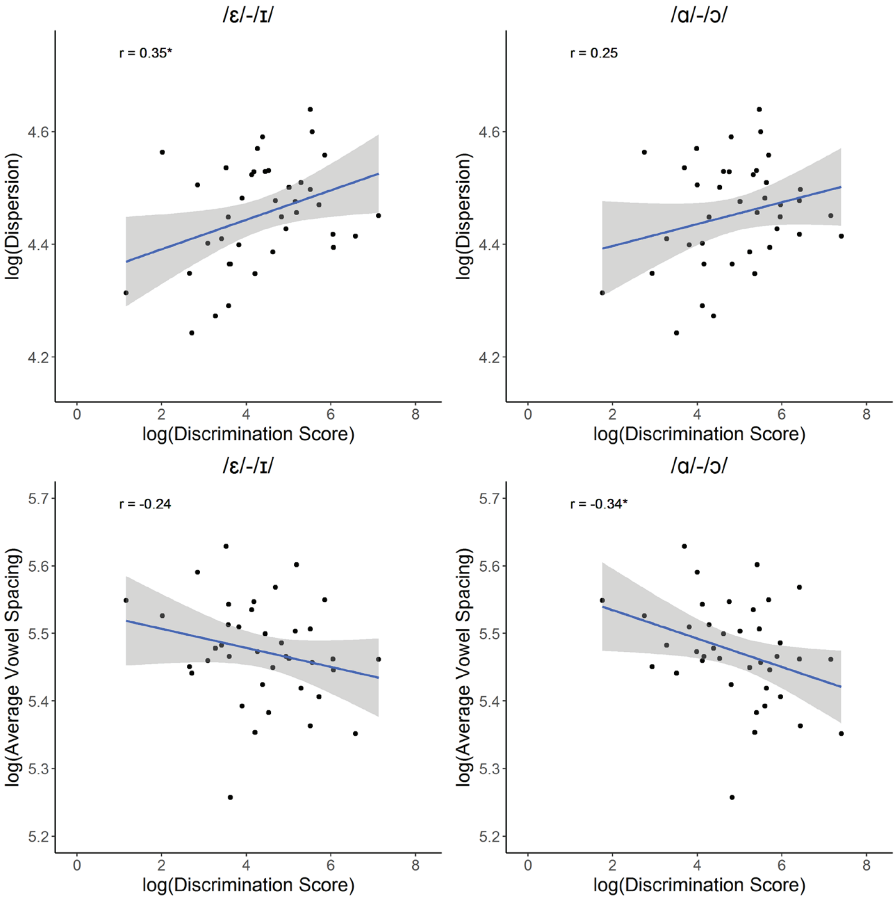

Fig. 2.6. Scatter plots of correlation analyses in MFCC space. The top row shows comparisons between discrimination score (x-axis) and within-phoneme variability (y-axis). The bottom row shows comparisons between discrimination score and average between-phoneme distance (y-axis). Left column shows results for the /ɛ/-/ɪ/ continuum, the right column for the /ɑ/-/ɔ/ continuum. Superimposed lines represent the best linear fit, shading represents 95% confidence interval.

general to be dependent on native language.

In addition to differences in language, and hence phoneme space, there were other differences between our study and Perkell et al., as we used different measures to quantify perception and production variability. In our discrimination task, we noticed a fairly high amount of variability between blocks within the same subject and the same continuum. This drove us to use the measure we called the Discrimination Score, which captured both the participants' best discrimination performance as well as their consistency across blocks. Perkell et al. did not report on the variability of discrimination performance within subjects, and simply used the measure of the participants' discrimination threshold. In terms of production measures, we have used measures that should take into account vowel distributions and perceptual warping of acoustic space to a larger degree. With respect to the measures in F1*F2 space, we used the area of the ellipse and Mahalanobis distance, whereas Perkell et al. used the standard deviation of the distribution and Euclidian distance. The measures used in our study, proposed earlier by Kartushina et al. (2013, 2014), take into account differential distribution shapes of the different phonemes, and therefore are likely to better reflect phoneme variability. Additionally, we characterized vowels in terms of mel-frequency cepstral coefficients, which imitate the transfer function of the cochlea in the human ear, thus capturing the vowels' acoustics in a way similar to the human ear. Although all these differences between the current study and the study by Perkell et al. may have contributed to the differences in effect size, note that using the same methods as Perkell et al. on the current data did not yield larger effect sizes.

The within-phonemic variability as measured by our Vowel Dispersion metric may capture both speech target precision as well as variability due to coarticulation across consonantal contexts. In order to estimate the effect of coarticulatory variability, we recalculated the dispersion measure after having removed the variance due to phonological context. Correlation analyses with discrimination scores led to similar results (*r*(36) = 0.23 for /ɛ/-/ɪ/, where it was 0.24 and *r*(36) = 0.26 for /ɑ/-/ɔ/, where it was 0.37). Thus for the /ɑ/-/ɔ/ continuum, it seems at least part of the association may be driven by coarticulatory variability. This would suggest that at least for this continuum, better discriminators show less coarticulatory variability. If we adopt a view on coarticulation where the phonological context affects an underlying phonemic target (Farnetani & Recasens, 2010), this is still in line with the hypothesis that the underlying target region is more precise (or more robust) for better discriminators. A more robust underlying target region would then leave less room for coarticulation effects to take place, and thus less coarticulatory variability.

The overall consistency of our results with those of Perkell et al. (2004, 2008) nevertheless shows that people with better auditory acuity have reduced production variability. These findings are, in turn, consistent with several recent models of speech production. Many of these models consider the goal of speech production to be at least partially an acoustic goal. Therefore, individual differences in auditory perception may well affect variability in speech production targets. One example of these models is presented in Perkell (2007, 2012), where speech production targets are explicitly considered to be regions in auditory space. According to this view, better auditory discrimination performance corresponds to having a higher resolution in auditory space, which in turn leads to more precise auditory speech production targets and therefore to more precise or less variable speech production.

Another important component of recent theoretical frameworks is the interaction between feedforward and feedback control of speech production. In the feedback control part of the system, the auditory target is activated during speech production, thus enabling comparison with incoming auditory feedback. When mismatches occur between predicted and actual auditory information, corrections can be implemented in real time and, over time, the feedforward mechanisms guiding motor targets can be updated and maintained. There are multiple possible means by which auditory acuity might influence this learning. First, under this sort of model (e.g., Tourville & Guenther, 2011), individuals who are better at discriminating speech sounds would then become better at detecting mismatches between feedback and speech targets, and would therefore update their feedforward mechanisms more readily. The end result of this process playing out over time is that speech productions at the periphery of the target region (in auditory space) would be recognized as an error for some individuals, but not for others. If it is recognized as an error, this may lead, over time, to changes in the feedforward commands as such 'errors' should be avoided, effectively decreasing the variability in speech production. Consistent with this mechanism, Villacorta et al. (2007) demonstrated that speakers with higher auditory acuity show a greater behavioral response to altered auditory feedback. Thus, a second possibility is that people with better auditory acuity respond more strongly to altered auditory feedback, thus decreasing the variability in their speech production over time. Finally, with respect to inter-phonemic distances, some studies have suggested previously that mismatches that bring the speech sound closer to a neighboring phoneme are more readily perceived or are compensated for more strongly than mismatches that bring the result farther away from a neighboring phoneme (Lametti, Krol, Shiller, & Ostry, 2014; Niziolek & Guenther, 2013). Over time, this may lead individuals who are more sensitive to these mismatches to produce speech sounds that are spaced further apart in auditory space, which in this study was quantified as larger average vowel spacing. Such a result was attained through simulations with the DIVA model (Tourville & Guenther, 2011; Perkell, 2012).

In this study we investigated the association between speech production and speech perception and whether this association would be dependent on local phoneme density. The results show that overall, speakers with higher auditory acuity produced vowels more distinctively, that is, vowels that were spaced further apart and with less within-category variability. This association did not depend on local phonemic density in vowel space. These findings corroborate current thinking about feedback processing during speech production and the role of auditory information. Furthermore, this study offers insights into individual variability in speech production, which to date is still not well understood. More specifically, our findings are consistent with predictions from current theoretical models of speech motor control, and suggest that speakers with higher auditory acuity have more precise speech production targets, which subsequently shapes their speech production.

APPENDIX

In addition to the results reported in the main text, and in order to be able to compare our results to the previous literature, the data were also analyzed with the methods reported by Perkell et al. (2008).

For the perception metric, the discrimination threshold was estimated for every block, as described in the main text. Subsequently, auditory acuity was estimated as the inverse of the discrimination threshold, and was averaged across blocks for each participant. With respect to the production data, all formant values were converted to mels using the following formula (Boersma & Weenink, 2013):

$$m = 1127 \cdot log(1+f/700)$$

Here, f is the formant estimate in Hertz and m is the estimate in mels. Dispersion was calculated for each phoneme as the average Euclidian distance to the centroid in F1-F2 space, and subsequently averaged across phonemes. Average vowel spacing was defined as the using Euclidian distance in mel between the centroids of all possible vowel pairs, and subsequently these values were averaged across vowel pairs.

Correlation tests reported no significant correlation between acuity and dispersion (r = -0.14), nor between acuity and average vowel spacing (r = 0.20). Note though, that the direction of the trend corresponds to the results reported in the main text (the sign of the correlation coefficients is different, as the acuity is the inverse of the discrimination threshold).

REFERENCES

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America, 116*(3), 1729–1738. doi: 10.1121/1.1779271

Behroozmand, R., Ibrahim, N., Korzyukov, O., Robin, D. A., & Larson, C. R. (2015). Functional role of delta and theta band oscillations for auditory feedback processing during vocal pitch motor control. *Frontiers in Neuroscience*, 9, 109. doi:10.3389/fnins.2015.00109

Behroozmand, R., Karvelis, L., Liu, H., & Larson, C. R. (2009). Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. *Clinical Neurophysiology, 120*(7), 1303–1312. doi:http://dx.doi.org/10.1016/j.clinph.2009.04.022

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*(6), 3153–3161. doi:10.1121/1.423073

Christoffels, I. K., van de Ven, V., Waldorp, L. J., Formisano, E., & Schiller, N. O. (2011). The Sensory Consequences of Speaking: Parametric Neural Cancellation during Speech in Auditory Cortex. *Plos One, 6*(5). doi:ARTN e18307 DOI 10.1371/journal.pone.0018307

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357-366. doi: 10.1109/TASSP.1980.1163420

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America, 70*(1), 45. doi:10.1121/1.386580

Farnetani, E., & Recasens, D. (2010). Coarticulation and connected speech processes. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second Ed., pp. 316-352). Oxford, UK: Blackwell Publishing Ltd. doi:10.1002/9781444317251.ch9

Fairbanks, G., & Guttman, N. (1958). Effects of Delayed Auditory-Feedback Upon Articulation. *Journal of Speech and Hearing Research, 1*(1), 12–22. Retrieved from <Go to ISI>://WOS:A1958CKC3600002

Franken, M. K., Hagoort, P., & Acheson, D. J. (2015). Modulations of the auditory M100 in an imitation task. *Brain and Language, 142*, 18–23. doi:10.1016/j.bandl.2015.01.001

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*(3), 363–376. doi: 10.3758/Bf03194885

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (2nd editio.). Wiley: Hoboken, NJ.

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport, 17*(13), 1375–1379. doi:10.1097/01.wnr.0000233102.43526.e9

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron, 69*(3), 407–422. doi:10.1016/j.neuron.2011.01.019

Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099. doi:10.1121/1.411872

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9469813

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*(28). doi:10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience, 14*(8), 1125–1138. doi:10.1162/089892902760807140

Kaernbach, C. (1991). Simple Adaptive Testing with the Weighted up-down Method. *Perception & Psychophysics, 49*(3), 227–229. doi: 10.3758/Bf03214307

Kartushina, N., & Frauenfelder, U. H. (2013). Foreign accents and native sloppiness : The role of individual native production on non-native vowel pronunciation. In *Phonetics, Phonology and Language Contact 13* (pp. 25–28).

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, 5(November), 1246. doi:10.3389/fpsyg.2014.01246

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107. doi:10.3758/BF03212211

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1493), 979–1000. doi:10.1098/rstb.2007.2154

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014). Brief periods of auditory perceptual training can determine the sensory targets of speech motor learning. *Psychological Science, 25*(7), 1325–36. doi:10.1177/0956797614529978

Lane, H., & Webster, J. W. (1991). Speech Deterioration in Postlingually Deafened Adults. *Journal of the Acoustical Society of America, 89*(2), 859–866. doi: 10.1121/1.1894647

Levitt, H. (1971). Transformed up-down Methods in Psychoacoustics. *Journal of the Acoustical Society of America, 49*(2), 467–&. doi: 10.1121/1.1912375

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *Journal of the Acoustical Society of America, 113*(5), 2850–2860. doi: 10.1121/1.1567280

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience, 33*(29), 12090–12098. doi: 10.1523/Jneurosci.1008-13.2013

Parkinson, A. L., Korzyukov, O., Larson, C. R., Litvak, V., & Robin, D. A. (2013). Modulation of effective connectivity during vocalization with perturbed auditory feedback. *Neuropsychologia, 51*(8), 1471–80. doi:10.1016/j.neuropsychologia.2013.05.002

Perkell, J. S. (2007). Sensory goals and control mechanisms for phonemic articulations. In *Proceedings of the 16th International Congress of the Phonetic Sciences* (pp. 169-174). Saarbruecken, Germany.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics, 25*(5), 382–407. doi:10.1016/j.jneuroling.2010.02.011

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contracts in related to their discrimination of the contrasts. *Journal of the Acoustical Society of America, 116*(4), 2338–2344. doi: 10.1121/1.1787524

Perkell, J. S., Lane, H., Ghosh, S. S., Matthies, M. L., Tiede, M., Guenther, F. H., & Ménard, L. (2008). Mechanisms of Vowel Production: Auditory Goals and Speaker Acuity. In *8th International Seminar on Speech Production* (pp. 29–32). Strasbourg, France.

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America, 120*(2), 966. doi:10.1121/1.2217714

Reiterer, S. M., Hu, X. C., Sumathi, T. A., & Singh, N. C. (2013). Are you a good mimic? Neuro-acoustic signatures for speech imitation ability. *Frontiers in Psychology*, 4. doi:Artn 782 Doi 10.3389/Fpsyg.2013.00782

Simpson, A. P. (2001). Dynamic consequences of differences in male and female vocal tract dimensions.

Journal of the Acoustical Society of America, 109(5), 2153. doi:10.1121/1.1356020

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621. doi: 10.1111/j.1749-818X.2009.00125.x

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981. doi:10.1080/01690960903498424

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America, 122*(4), 2306–2319. doi: 10.1121/1.2773966

3

FEEDBACK CONSISTENCY

ABSTRACT

Previous research on the effect of perturbed auditory feedback in speech production has focused on two types of responses. On the one hand, speakers issue quick compensatory motor commands in response to unexpected perturbations. On the other hand, speakers adapt feedforward motor programs in response to feedback perturbations, in order to avoid future errors. The current study investigates the relation between these two types of responses to altered auditory feedback. Specifically, it is hypothesized that the previous history of feedback perturbations (its consistency) may determine whether speakers adapt their feedforward motor programs. In an altered auditory feedback paradigm, formant perturbations were either consistent across trials, or it was unpredictable whether feedback on a given trial would be perturbed or not. The results show that speakers' responses were indeed affected by feedback consistency, with stronger speech changes in the consistent condition compared to the inconsistent condition. This suggests that the speech production system keeps track of the feedback channel's consistency. Finally, it is discussed that present models could account for this result if they are updated to include a mechanism to keep track of feedback consistency and use it to modulate the adaptation of forward models.

3.1 INTRODUCTION

Speaking is a remarkable human motor skill. It requires very precise control and coordination of various muscles spread throughout the vocal tract, in order to produce an acoustic sequence that allows a listener to reconstruct the intended message. In addition, there are several sources of nuisance that can potentially distort accurate transmission: environmental noise, speaker idiosyncrasies, variability of accents and dialects, etc. Yet, somehow, we accomplish this difficult task many times every day, seemingly without effort.

One way in which the speech system can deal with disturbances and changes in the environment is through monitoring sensory feedback. It is well established that the speech production system interacts in a complex way with sensory feedback. While auditory feedback may not be strictly necessary for speech production, incorrect or unexpected feedback leads to disturbances in speech production (Elman, 1981; Fairbanks & Guttman, 1958). This has been shown with delayed auditory feedback (Fairbanks & Guttman, 1958), as well as with altered auditory feedback, where pitch (Burnett, Senner, & Larson, 1997; Elman, 1981), amplitude (Bauer, Mittal, Larson, & Hain, 2006), formant frequency (Houde & Jordan, 1998; Purcell & Munhall, 2006b), or fricative noise (Casserly, 2011) was manipulated.

These studies, where researchers manipulated speakers' auditory feedback during the act of speaking, show that the speech production system uses auditory feedback to inform and control the speech production process. Broadly, research has focused on auditory feedback being used on two different time-scales. In the short term, unexpected auditory feedback may lead to immediate corrective responses within the same trial (Burnett et al., 1997; Purcell & Munhall, 2006b). This line of research has shown that speakers on average quickly compensate for sudden auditory feedback perturbations, although sometimes the perturbation is followed (i.e., instead of compensating for the change, the speaker makes adjustments in the same direction as the change; Hain et al., 2000). In the longer term, speakers show evidence of adaptation to consistent feedback (Houde & Jordan, 1998, 2002; Jones & Munhall, 2000). This is usually described as an aftereffect, which is a lingering adaptation after the perturbation has been removed. The presence of aftereffects show that, over time, speakers have adapted to the new sensorimotor environment by changing their feedforward speech motor commands accordingly (Purcell & Munhall, 2006a).

Theoretical frameworks have been developed to account for feedback-based speech adjustments. For example, in the DIVA model (Guenther & Vladusich, 2012; Guenther, 2006), a distinction is made between the feedforward and the feedback control systems. The feedback control system compares the expected sensory consequences to the observed sensory input (i.e., the feedback). A mismatch leads to corrective behavioral adjustments. In addition, over time these adjustments could be incorporated in the feedforward system in order to avoid future errors. A similar framework is the state feedback control (SFC) model (Houde, Kort, Niziolek, Chang, & Nagarajan, 2013; Houde & Nagarajan, 2011), where auditory feedback is used to control and update an internal estimate of the dynamic state of the speech production system. In this model as well, unexpected auditory feedback perturbations lead to compensatory behavioral adjustments and/or longer-term changes to the internal forward model.

An open question is how the short-term and long-term feedback-based speech adjustments relate to each other. In the remainder of this article, we will refer to the immediate response to altered auditory feedback as compensation, and to longer-term changes in feedforward commands, as evidenced by aftereffects, as adaptation. So how does the speech system decide when to adapt its feedforward commands? As theorized already in Bayesian approaches to sensorimotor learning and adaptation (Franklin & Wolpert, 2011), the usefulness of adaptation may vary across contexts. If a mismatch between expected and observed feedback is consistent time after time, it would make sense to adapt feedforward commands to avoid the prediction error in future attempts. However, sometimes the feedback perturbation may be an isolated event, and so adaptation would lead to an additional error on the next attempt. This leads to the hypothesis that speech motor adaptations will depend on the context in which the feedback is perturbed. If the perturbation is consistent and/or predictable, we predict adaptation, whereas there should be no adaptation when the feedback perturbation is inconsistent and/or unpredictable. Testing this hypothesis is the main goal of the present study.

In addition, while many studies have investigated compensation or adaptation, only few have looked at them together, which makes it possible to study their interaction. In an altered auditory feedback paradigm with on-line adjustments to the first formant of the vowel being produced, we tried to disentangle the speaker's (short-term) response to unexpected feedback perturbation (compensation) and the speaker's adaptation to consistent feedback perturbations over time, by comparing an inconsistent condition where the type of auditory feedback (perturbed or unperturbed) is unpredictable with a consistent condition where the feedback perturbation is consistent. We expected that speakers would show compensation by changing their speech output in the perturbed trials in both conditions (vs. the unperturbed baseline trials). In addition, we expected speakers to show adaptation in the form of an aftereffect in the consistent condition, but not (or less so) in the inconsistent condition. Comparing perturbed trials between inconsistent and consistent conditions, we expect more change in the consistent condition, as this condition would show both compensation and adaptation.

3.2 MATERIALS & METHODS

3.2.1 Participants

Twenty-six healthy volunteers (age: M = 22, SD = 2.7; 17 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local Ethics Committee of the Social Sciences faculty of Radboud University. All participants had normal hearing, were native speakers of Dutch and had no history of speech and/or language pathology. Two participants were excluded because their speech production was too quiet to trigger feedback perturbation.

3.2.2 Paradigm

The experiment consisted of four blocks of 80 trials each (see Figure 3.1). On every trial, participants were instructed to vocalize /e/ as soon as the letters <ee> appeared on a computer screen and keep doing so until the letters disappeared (3 seconds later). During vocalization, participants' speech was recorded and fed back to them as auditory feedback through headphones. Depending on the experimental condition, in some trials (perturbation trials) the auditory feedback was manipulated by shifting the first formant (F1) by 6.7% up or down. There were two conditions. In the consistent condition, a block started with 20 start trials (where auditory feedback was not perturbed), followed by 40 perturbed trials (where all 40 trials where perturbed by shifting F1 in the same direction) and finally 20 end trials (where auditory feedback was unperturbed). In the inconsistent condition, the block also started with 20 start trials and ended with 20 end trials, but the 40 trials in the middle were randomly assigned to be either a non-perturbed or a perturbed trial. Furthermore, in a given block in the inconsistent condition there were always 20 perturbed trials in total, trial 21 was always perturbed, and all perturbed trials were perturbed in the same manner (same direction and magnitude). Each condition (consistent and inconsistent) was provided to every participant twice, with perturbations being once an upward F1 shift and once a downward F1 shift, amounting to four blocks in total (see Figure 3.1). The order of the blocks was counterbalanced while making sure an inconsistent block never followed or preceded another inconsistent block (and similarly for consistent blocks).

All voice recordings were made on one channel using a Sennheiser ME64 cardioid microphone, which was set up in a soundproof booth and connected to a dedicated soundcard Motu MicroBook II outside the booth, which was in turn connected to a Windows laptop. Auditory feedback was delivered through the same soundcard which was also connected to Sennheiser HD 2801-13 headphones. Stimulus presentation and sound recording times

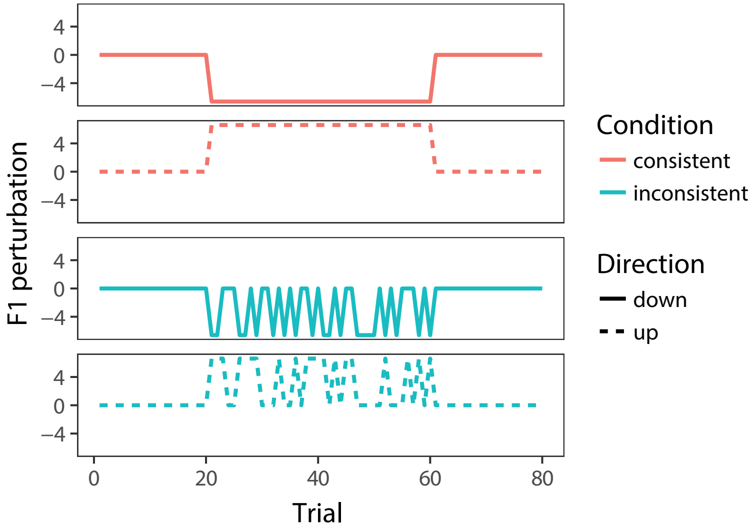

Fig. 3.1. Overview of the experimental blocks, in a 2-by-2 design (consistency by perturbation direction). Order of the blocks was counterbalanced across participants, so that two blocks of the same consistency never followed each other. The sequence of perturbed and unperturbed trials in the inconsistent condition is an example; trial-type order was randomized across participants (see main text for constraints on randomization).

were controlled by the same Windows laptop running Audapter (Cai, Boucek, Ghosh, Guenther, & Perkell, 2008; Tourville, Cai, & Guenther, 2013) and MathWorks Matlab (R2013b).

3.2.3 Analysis

For every trial, the recording of the participants' speech recording was marked for speech onset and offset by visual inspection, and F1 was estimated in Mels (Stevens, Volkmann, & Newman, 1937). For an analysis of the mean F1 across trials, the average F1 value was calculated from 50ms until 1500ms after speech onset. Trials where formant estimation failed within this window were rejected (across subjects, on average 1.5 (sd = 4.6) trials were removed from further analysis).

Further analyses were carried out with R (R Core Team, 2013). For every subject and every block, F1 values were normalized in the following way.

$$F1_{norm}=-d\cdot(F1- F1_{start})/F1_{start}$$

Here, $F1_{start}$ is the average F1 value of the start trials for that subject in that block, and *d* is the sign of the direction of perturbation in that particular block. This leads to $F1_{norm}$ values being expressed as the percentage change in the opposite direction to the perturbation direction (irrespective of which direction that was) relative to the average F1 in the start phase (first 20 trials of the block). In other words, if participants compensate for the F1 perturbation, we expect $F1_{norm}$ to be positive, irrespective of the perturbation direction.

Statistical testing was done by means of linear mixed effects models as implemented by the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015). In a first step, an appropriate model was selected by means of Akaike's Information Criterion (AIC). A series of models that differed with respect to their random effects structure were compared. Subsequently, the most appropriate random effects structure was selected and models with varying fixed effect structures were compared. The reported p values were calculated using a Satterthwaite approximation of the degrees of freedom.

In a second analysis, we took the temporal development of the F1 contour within a trial into account by calculating the average F1 contour for every participant and every trial type. For statistical inference, a non-parametric permutation test was performed with a clustering method to correct for multiple comparisons (Maris & Oostenveld, 2007), as implemented in the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). This was done for the data between 50ms-1500ms after speech onset to determine whether there was a difference between the perturbed and unperturbed trials within the inconsistent condition. Samples for which the contrast Perturbed - Unperturbed exceeded an uncorrected α level of .05 were temporally clustered. Cluster-level statistics were calculated by summing the t-statistics. Next, a permutation distribution of statistics was calculated by randomly exchanging data between the two trial types, and calculating the maximal positive and negative cluster-level statistics for every permutation (for a total of 10,000 permutations). The observed cluster-level statistic was tested against the permutation distribution.

3.3 RESULTS

In a first set of analyses, the average normalized F1 values were compared across conditions. An overview is presented in Figure 3.2, where normalized F1 is shown as a function of Trial Type in both Conditions (random vs. consistent). As the figure suggests, participants altered the F1 in their speech output as a function of Condition and Trial Type.

The data from the perturbed and end trials was entered in a linear mixed effects model. Model selection was done in two steps. First, a full fixed effects structure was selected (main effects of Condition and Trial Type, as well as their interaction), while varying the random effects in four different models. Table 3.1 shows the model comparison. It can be concluded from the table that the model with the maximal random effects structure (by-participant random slopes for Condition, Trial Type and their interaction) fits the data best. Subsequently, three models were compared while keeping the random effects structure maximal. From Table 3.2 it is clear that while the model that included both Condition and Trial Type main effects was a better fit to the data then the model with only a fixed intercept, the model including the interaction did not do better. Therefore, the model without interaction was selected and its output is shown in Table 3.3. As there was no fixed intercept included in the model, a significant main effect indicates the main effect's coefficient is significantly different from zero, and therefore F1 in this condition different from the average start F1 (given the normalization). Table 3.3 shows that F1 was indeed significantly altered in perturbed trials in the consistent condition, but not in the inconsistent condition. In addition,

Table 3.1. Model random effect structure comparison

Model random effect structure	df	AIC	X^2	p
(1 | Participant)	6	-10275		
(1 + TrialType | Participant)	8	-10371	99.53	<.001*
(1 + Condition | Participant)	8	-10598	227.78	<.001*
(1 + Condition*TrialType | Participant)	15	-10747	162.08	<.001*

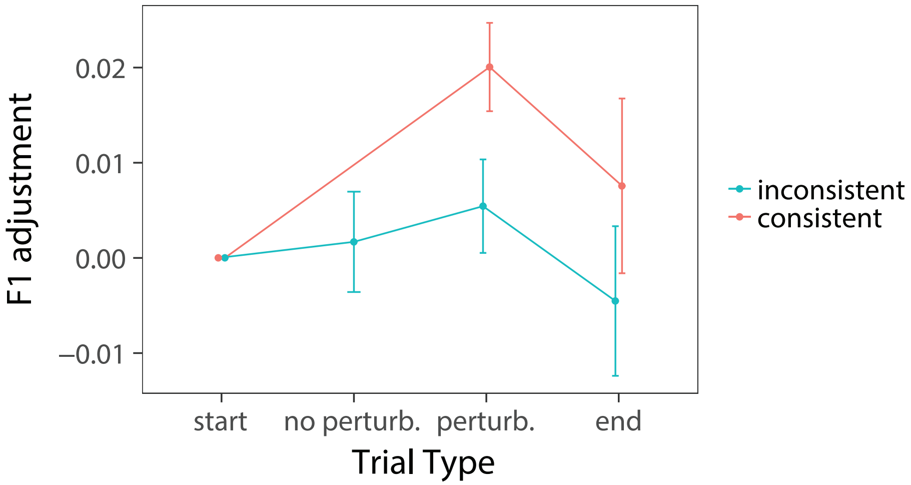

Fig. 3.2. F1 adjustments as a function of trial type and consistency. Error bars reflect standard errors across participants. (start = unperturbed trials at the beginning of a block; perturb. = perturbation trials; no perturb. = unperturbed trials in the middle portion of an inconsistent block, end = unperturbed trials at the end of a block)

F1 in the end trials was lower than in the perturbed trials (p = 0.022). This suggests that under altered auditory feedback, participants responded by shifting their F1 in the opposite direction compared to the perturbation, but only or especially under consistent perturbation across trials. After auditory feedback returned to normal, F1 returned to baseline.

In order to compare the perturbed and unperturbed trials in the middle section of the inconsistent blocks, a second set of analyses was carried out on the inconsistent condition only. Tables 3.4 and 3.5 show the results for the model selection procedure. Table 3.4 shows that adding by-participant random slopes for Trial Type significantly improved the model. In other words, the relationship between average F1 and Trial Type varied across participants. As can be seen from Table 3.5, including a fixed main effect for Trial Type did not significantly improve the model's fit to the data, suggesting the average normed F1 did not differ as a function of altered auditory feedback, although Figure 3.2 suggests trends in the expected direction.

As perturbed and unperturbed trials are randomly mixed in the inconsistent condition, it is expected that responses in the inconsistent

Table 3.2. Model fixed effect structure comparison

Model	df	AIC	X^2	p
1 + (Condition*TrialType | Participant)	12	-10777		
0 + Condition + TrialType + (…)	14	-10781	8.58	0.014*
0 + Condition * TrialType + (…)	15	-10779	0.10	0.75

Table 3.3. Model: F1 ~ 0 + Condition + TrialType + (1 + Condition*TrialType | Participant)

Fixed effects	Estimate	SE	t	p (Satt.)
Condition.Inconsistent	.0075	.0063	1.20	.24
Condition.Consistent	.026	.0058	4.51	< .001*
TrialType.End	-.014	.0058	-2.46	.022*

Table 3.4. Model random effect structure comparison

Model random effect structure	df	AIC	X^2	p
(1 | participant)	5	-6632.0		
(1 + TrialType | participant)	10	-6681.1	59.1	<.001*

Table 3.5. Model fixed effect structure comparison

Model	df	AIC	X^2	p
1 + (1 + TrialType | participant)	8	-6707.2		
0 + TrialType + (…)	10	-6707.3	4.13	0.13

perturbed trials will take some time, whereas in the consistent condition participants can already expect the perturbation (and have built up adaptation over the course of previous trials) and therefore adapted their F1 production from the start of the trial. Therefore, to investigate the difference between perturbed and unperturbed trials in the inconsistent condition closer, the time course of the F1 contour in the inconsistent condition was analyzed. Figure 3.3 shows the F1 contour (or the difference in F1 contour from the average contour in the start trials) as a function of condition and direction of perturbation. For illustrative purposes, the data from the consistent condition are provided as well, although they were not further analyzed.

Interestingly, the F1 contour for perturbed trials in the inconsistent condition overlaps with that for the unperturbed trials at the beginning of the trial, but the two contours diverge towards the end of the trial. A cluster-based permutation test was carried out to determine whether the perturbed trial data differed from the unperturbed trial data across the time window 100ms-1500ms after speech onset. For the blocks with downward F1 perturbation, the F1 in the unperturbed trials was lower than the F1 in the perturbed trials (one-sided test; *p* = 0.0042, CI = [0.0029 0.0055]), whereas it was higher in the positive F1 perturbation blocks (one-sided test; *p* = 0.043, CI = [0.039 0.047]). These effects are mainly driven by a single large cluster starting from 666ms after speech onset for the downward F1 shift direction (until the end of the time window), and from 1162ms until 1448ms for the positive F1 shift direction. The time course of the uncorrected t statistics in Figure 3.4 suggests that F1 change indeed increases with time.

Finally, it was examined whether the difference in F1 adjustment between the consistent and inconsistent conditions (consistency-related adjustment) for the perturbation trials was due to compensation (immediate feedback

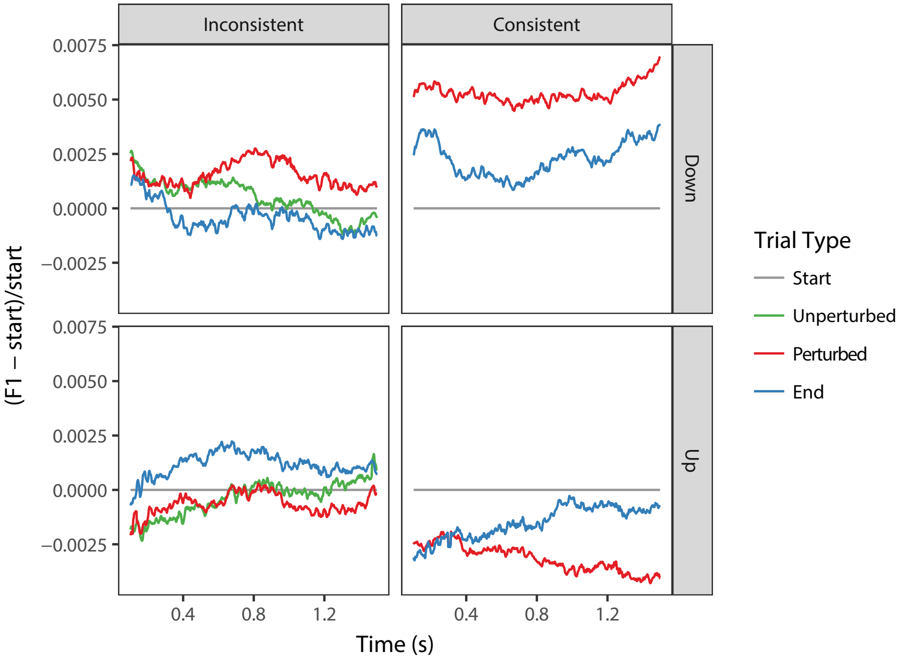

Fig. 3.3. F1 time courses from 0.1 to 1.5s after speech onset, as a function of trial type (color), consistency (column), and perturbation direction (row).

responses) or adaptation (altered feedforward commands). To that end, a linear regression model was calculated where the consistency-related adjustment (across the trial) was regressed against the amount of compensation and the amount of adaptation for each participant. The compensation values were quantified by taking the difference in F1 adjustments between inconsistent perturbation and non-perturbation trials for that participant (between 1.1s and 1.5s after speech onset), and the adaptation values were quantified by taking the average F1 adjustment for that participants' first 10 trials in the end phases of the sustained condition (between 0.1s and 0.4s after speech onset). Table 3.6 shows the results of this linear regression model. It can be seen that part of the consistency difference can be significantly explained by adaptation, suggesting the consistent perturbation leads to more adaptation. There is no significant relation with compensation. In addition, there was no significant association between adaptation and compensation (*r*(46) = -0.054, *p* = 0.72). Overall, this suggests that compensation and adaptation are two distinct processes, with differences between responses in consistent and inconsistent conditions mainly due to adaptation.

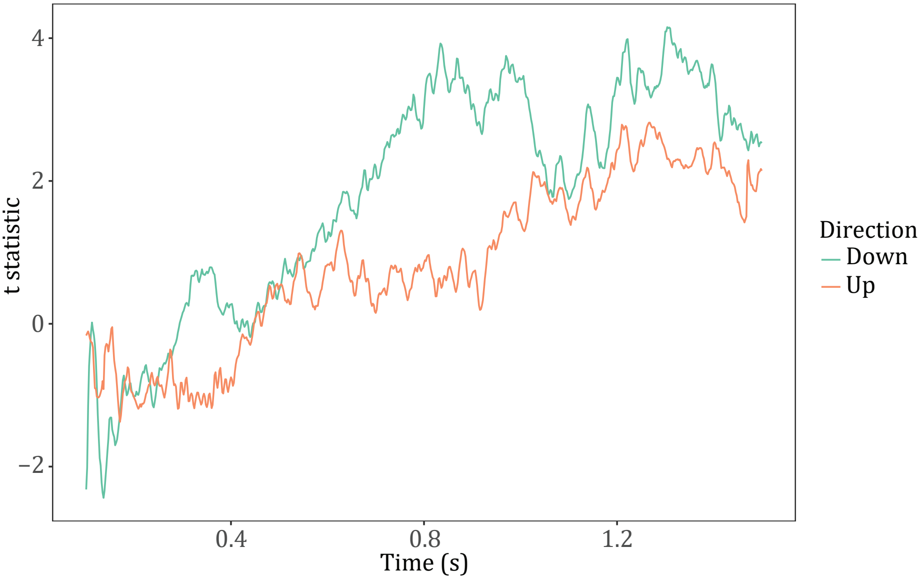

Fig. 3.4. T statistic (uncorrected) for the contrast inconsistent perturbed vs. non-perturbed trials across time. Colors indicate the perturbation direction.

Table 3.6.

	Estimate	SE	t	p
(Intercept)	.0021	.0015	1.41	.16
Adaptation	.29	.13	2.25	**.030***
Compensation	.014	.065	.21	.84
Adaptation:Compensation	-.68	7.16	-.095	.92

3.4 DISCUSSION

The current study used an altered auditory feedback paradigm to investigate how feedback-related speech responses are affected by the recent history of feedback perturbations. The results indicate that the consistency of this history indeed affects how speakers respond to altered auditory feedback, suggesting that more consistent feedback-based prediction errors lead to stronger behavioral adjustments.

Specifically, there was a significant difference between inconsistent and consistent conditions in the perturbed trials. The perturbed trials in both conditions were exactly the same, so any response difference must be due to the context of the trial (whether it was embedded in an inconsistent or consistent feedback environment). This suggests that speakers' motor adjustments are indeed affected by the history of previous trials' perturbations, as hypothesized. This could reveal a general build-up of adaptation across consistent trials, or, alternatively, an adjustment of the gain of feedback-driven error processing mechanisms. In an inconsistent environment, the feedback channel is less consistent and could therefore could be considered less reliable, leading to a reduction in the gain on this feedback channel. In the consistent condition, on the other hand, the increased reliability would lead to a higher gain on feedback-related processing. The idea of modulating the gains on feedback processing is also consistent with findings from the broader (non-speech) motor control literature (Gonzalez Castro, Hadjiosif, Hemphill, & Smith, 2014). These authors showed in an arm reaching task with force field perturbations that the rate of trial-to-trial adaptations was associated with consistency of the environment, showing quicker/stronger adaptation with more consistent perturbations.

In the current study, there was no significant aftereffect in the overall analysis, although Figure 3.2 suggests a trend in the expected direction. Previously, the presence of an aftereffect has commonly been taken as evidence in favor of adaptation of feedforward commands. The lack of strong evidence for adaptation of the feedforward commands in the current study may be due to the rather small magnitude of our perturbation, and/or to the long trial length. Longer trials may provide more opportunity for the adaptation effect to disappear in the end trials. Nevertheless, the association we find between participants' adaptation and the consistency-related difference between conditions on the perturbation trials suggests that consistency leads to adaptation.

With respect to compensation, an analysis of the F1 time course did find a difference between perturbed and non-perturbed trials in the inconsistent condition, suggesting that speakers showed compensation or within-trial feedback responses. The analysis of the F1 time courses suggested in addition (1) that compensation increases over time, as expected, and (2) that speakers adjusted their F1 in response to altered auditory feedback more strongly for downward perturbations compared to upward F1 shifts. We speculate this effect of perturbation direction may be due to asymmetry in the Dutch vowel space: starting from /e/, there are more close-by vowel phonemes with decreasing F1 compared to increasing F1. As Niziolek & Guenther (2013) have shown, when auditory feedback is manipulated by shifting it towards a close-by phoneme boundary, speakers tend to respond more strongly in order to avoid misinterpretation of the vowel identity by the listener.

Several models have been proposed that can account for speech adjustments under altered auditory feedback. The DIVA model (Guenther, Ghosh, & Tourville, 2006; Guenther, Hampson, & Johnson, 1998; Guenther & Vladusich, 2012; Tourville & Guenther, 2011) makes a clear distinction between feedforward and feedback control subsystems. Immediate feedback responses or compensation is generated by the feedback control subsystem, the output of which is integrated with the feedforward commands by the motor cortex. Adaptation of the feedforward control subsystem occurs subsequently, as the feedback-based corrections are integrated in the feedforward commands to avoid similar errors on subsequent trials. We speculate that the DIVA might account for the present results by modulating its weights on the feedback control signal (Guenther et al., 2006; Tourville, Reilly, & Guenther, 2008). However, the model is not specific about how these weights could be modulated by the system. In addition, the DIVA model predicts that adaptation and compensation are associated: feedforward motor control is adapted by integrating a weighted version of the compensation response of the previous trial. This hypothesis was also not borne out by the current results, as there was no correlation between compensation and adaptation. In fact, the results suggest compensation and adaptation are two separate processes, with mainly adaptation being affected by feedback consistency.

A somewhat different model that was inspired by previous modeling work in non-speech motor control is the state feedback control model (SFC, Houde et al., 2013; Houde, Niziolek, Kort, Agnew, & Nagarajan, 2014; Houde & Nagarajan, 2011). This model differs from DIVA in that it does not make a clear distinction between immediate feedback responses (compensation) and adaptations of the feedforward commands. Instead, prediction errors generated by unexpected sensory feedback are used to update an internal estimate of the dynamical state of the vocal tract, which in turn is used for generating the next motor commands. The model assumes a Kalman gain function on the feedback prediction error. The Kalman gain depends on the variability in the observed feedback, and thus can upregulate the influence of feedback when it is reliable (low variability), or downregulate when it is not. However, this gain controls the influence of feedback on the state estimate, and therefore on the next generation of motor commands. It is unclear how feedback prediction errors would modulate the feedforward commands (i.e., the mapping between speech targets and appropriate motor controls). Note that these authors have proposed more recently that there should be a degree of independence between short-term compensation and forward model adaptation, due to a possible dissociation between affected compensatory behavior and affected forward model adaptation in clinical disorders

(Houde, Nagarajan, Parrell, & Ramanarayanan, 2017). Overall, it seems that both DIVA and the state feedback model of speech motor control would have to be adapted in order to fully account for the present results. Although both models have the ability to control the gain on auditory feedback processing, an explicit way of modulating the gain based on environmental consistency and/or a way to control subsequent feedforward adaptations would have to be added to the models.

Overall, the present report suggests that speakers' feedback-based speech adjustments depend on the consistency of past feedback errors. This can be implemented in the speech system by keeping track of the feedback error history, or, as in a Kalman filter, by keeping track of the variability of the feedback signal. If a mismatch between expected and observed auditory input is consistent, it is advantageous to adapt feedforward control to this new environment. If it is sporadic, strong adaptation may in fact cause additional errors, and are therefore not warranted. The current data are in line with this view. In addition, we suggest current models need to be updated in order to fully account for these data. This can be done by including a mechanism to keep track of the variability or consistency of the feedback signal (e.g., a Kalman filter as in the SFC model), and by allowing this consistency to modulate the feedback-based adaptation of the feedforward commands. In addition, the current data suggest that short-term compensation and forward model adaptation are two distinct processes.

REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). doi:10.18637/jss.v067.i01

Bauer, J. J., Mittal, J., Larson, C. R., & Hain, T. C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: an automatic mechanism for stabilizing voice amplitude. *The Journal of the Acoustical Society of America, 119*, 2363–2371. doi:10.1121/1.2173513

Burnett, T. A., Senner, J. E., & Larson, C. R. (1997). Voice F0 responses to pitch-shifted auditory feedback: a preliminary study. *Journal of Voice, 11*(2), 202–211. doi:10.1016/S0892-1997(97)80079-3

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In *Proceedings of the 8th Intl. Seminar on Speech Production* (pp. 65–68). Strasbourg, France.

Casserly, E. D. (2011). Speaker compensation for local perturbation of fricative acoustic feedback. *The*

Journal of the Acoustical Society of America, 129, 2181–2190. doi:10.1121/1.3552883

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America, 70*(1), 45. doi:10.1121/1.386580

Fairbanks, G., & Guttman, N. (1958). Effects of Delayed Auditory-Feedback Upon Articulation. *Journal of Speech and Hearing Research, 1*(1), 12–22. Retrieved from <Go to ISI>://WOS:A1958CKC3600002

Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron, 72*(3), 425–442. doi:10.1016/j.neuron.2011.10.006

Gonzalez Castro, L. N., Hadjiosif, A. M., Hemphill, M. A., & Smith, M. A. (2014). Environmental consistency determines the rate of motor adaptation. *Current Biology : CB, 24*(10), 1050–61. doi:10.1016/j.cub.2014.03.049

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders, 39*(5), 350–365. doi:10.1016/j.jcomdis.2006.06.013

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301. doi:10.1016/j.bandl.2005.06.001

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105*(4), 611–633. Retrieved from <Go to ISI>://WOS:000076943900001

Guenther, F. H., & Vladusich, T. (2012). A Neural Theory of Speech Acquisition and Production. *Journal of Neurolinguistics, 25*(5), 408–422. doi:10.1016/j.jneuroling.2009.08.006

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research, 130*(2), 133–141. doi:10.1007/s002219900237

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9469813

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech Language and Hearing Research, 45*(2), 295–310. doi:Doi 10.1044/1092-4388(2002/023)

Houde, J. F., Kort, N. S., Niziolek, C. A., Chang, E. F., & Nagarajan, S. S. (2013). Neural evidence for state feedback control of speaking. In *Proceedings of Meetings on Acoustics* (Vol. 19, pp. 060178–060178). Acoustical Society of America. doi:10.1121/1.4799495

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5(28). doi:10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Parrell, B., & Ramanarayanan, V. (2017). Advances in modeling speech production as state feedback control. *Stem-, Spraak- En Taalpathologie: 7th International Conference on Speech Motor Control*, Abstracts, 22(Supplement), 7.

Houde, J. F., Niziolek, C. A., Kort, N. S., Agnew, Z., & Nagarajan, S. S. (2014). Simulating a state feedback model of speaking. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Muecke (Eds.), *Proceedings of the 10th International Seminar on Speech Production* (pp. 202–205). Cologne, Germany.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America, 108*(3), 1246. doi:10.1121/1.1288414

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. doi:DOI 10.1016/j.jneumeth.2007.03.024

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience, 33*(29), 12090–12098. doi:Doi 10.1523/Jneurosci.1008-13.2013

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). Fieldtrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience, 2011*(2011). doi:doi:10.1155/2011/156869

Purcell, D. W., & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America, 120*(2), 966. doi:10.1121/1.2217714

Purcell, D. W., & Munhall, K. G. (2006b). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America, 119*(4), 2288. doi:10.1121/1.2173514

R Core Team. (2013). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statictical Computing. Retrieved from http://www.r-project.org

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America, 8*(3), 185–190. doi:10.1121/1.1915893

Tourville, J. A., Cai, S., & Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech (pp. 060180–060180). doi:10.1121/1.4800684

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981. doi:10.1080/01690960903498424

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage, 39*(3), 1429–1443. doi:http://dx.doi.org/10.1016/j.neuroimage.2007.09.054

4

FOLLOWING AND OPPOSING RESPONSES IN SENSORIMOTOR ADAPTATION: WHY RESPONSES GO BOTH WAYS

ABSTRACT

When talking, speakers continuously monitor and use the auditory feedback of their own voice to control and inform speech production processes. When speakers are provided with auditory feedback that is perturbed in real time, most of them compensate for this by opposing the feedback perturbation. But some speakers follow the perturbation. In the current study we investigated whether the state of the speech system at perturbation onset may determine what type of response (opposing or following) is given. The results suggest that whether a perturbation-related response is opposing or following depends on ongoing fluctuations of the speech system: The motor system initially responds by doing the opposite of what it was doing. This effect and the non-trivial proportion of following responses suggest that current production models are inadequate: They need to account for why responses to unexpected sensory feedback depend on the production-system's state at the time of perturbation.

4.1 INTRODUCTION

An important aspect of action control is performance monitoring through sensory feedback. Such control allows us to either confirm an appropriate action plan, adapt to a changing environment, or learn from our mistakes. For example, when throwing a ball, visual feedback could show us that the throw was successful, or it could indicate the need for adaptation to a changed environment (e.g., if the wind suddenly shifts). Similar processes are at play for auditory feedback in speech or musical production.

The importance of auditory feedback during speech production is well established using the technique of altered auditory feedback (Burnett, Freedland, Larson, & Hain, 1998; Houde & Jordan, 1998). For example, Burnett et al. (1998) manipulated the pitch frequency of speakers' auditory feedback in real time. Typically, speakers respond by adjusting the pitch in their output in the opposite direction to the pitch shift (Burnett et al., 1998; Liu & Larson, 2007). In other words, speakers seem to compensate for unexpected changes in auditory feedback so that their actual output more closely matches their intended output. We argue here that there is more to sensorimotor adaptation than these opposing responses.

Vocal motor control is a noisy process, so needs constant feedback monitoring. In fact, in both speech (Akagi, Iwaki, & Minakawa, 1998) as well as singing (Akagi & Kitakaze, 2000), pitch does not remain constant at the target pitch level, but fluctuates around the target. Pitch fluctuations are in fact an identifying feature of human pitch control, and removing it makes synthesized speech or song sound more robotic or unnatural (Akagi et al., 1998). These fluctuations are maintained by constant feedback monitoring and subsequent updating of the vocal motor commands. Using a vowel production task, Niziolek et al. (2013), for instance, showed that the auditory feedback control system is indeed sensitive to very small deviations in vocal production.

Feedback monitoring and subsequent adaptation is accounted for in a number of theoretical frameworks of speech motor control (Guenther, Ghosh, & Tourville, 2006; Hickok, 2012; Houde & Nagarajan, 2011). These models hypothesize the existence of so-called internal forward models, which predict the sensory (e.g., auditory) consequences of actions in real time. This prediction is compared with the incoming auditory feedback in order to monitor for speech errors. A mismatch will drive the speech motor system to initiate corrective (i.e. compensating) motor commands.

In contrast to these model predictions, however, several studies have reported that sometimes, instead of feedback compensation, responses are observed that follow the direction of the altered feedback (Burnett et al., 1998; Hain et al., 2000; Larson et al., 2007). These following responses are less frequent than opposing (i.e., compensating) responses, and are usually reported at the subject level. However, looking at single trials, Behroozmand et al. (2012) showed that even subjects that show an opposing response on average may show following responses on some trials. Therefore, the focus on the average response may have obscured the field’s view on the nature of following responses.

The presence of following responses has led some authors to suggest that voice pitch control has two feedback modes: one for tracking an external referent (eliciting following responses) and another for correcting for internal disturbances (eliciting opposing responses; Burnett et al., 1998; Hain et al., 2000). For example, in the context of choir singing, one singer might follow the pitch of another, an external referent (e.g., go flatter if the fellow singer is singing flat). However, the feedback signal of one’s own voice should activate the feedback mode for internal disturbances and therefore lead to an opposing response (e.g., go sharper when you’re singing flat). Both feedback modes may thus be simultaneously active.

The small number of studies that have looked at following responses suggested that such responses occur more often when the pitch manipulation is larger and that following responses often have a shorter duration than opposing responses (Burnett et al., 1998). Behroozmand et al. (2012) showed that predictable altered feedback may encourage a tendency to imitate the feedback when it changes in the same way over and over again.

In the current study, we investigated what factors play a role in feedback-based pitch control. A paradigm was used where participants tried to match a pitch target while vocalizing. Participants received auditory feedback through headphones, which sometimes was unexpectedly pitch-shifted for 500ms. None of the participants were aware of the pitch manipulations. We expected participants on average to compensate for the feedback, but at the single trial level to sometimes follow and sometimes oppose the pitch shift. Opposing/following balance may depend not only on whether the perturbation is internal or external, or on how large the perturbation is, but also on the state of the system at the time of the perturbation. We therefore explored the possibility that there are constraints on the action control process that limit how the system can respond to a perturbation. If so, then the current pitch fluctuation should be predictive of the kind of response.

4.2 METHODS & MATERIALS

4.2.1 Participants

Thirty-nine healthy volunteers (age: M = 22, SD = 3.6; 27 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local ethics committee (CMO region Arnhem / Nijmegen). All participants had normal hearing, were native speakers of Dutch and had no history of speech and/or language pathology. The sample size was based on a power analysis of MEG connectivity effects in Ford et al. (2005), indicating one would need about 33 subjects (dz = 0,506; power = 80%). We therefore planned to test at least 36 participants. This sample size exceeds that in most related previous studies (e.g., Behroozmand et al., 2012).

4.2.2 Paradigm

Participants performed a tone-matching task while their brain activity was measured using magnetoencephalography (MEG). The MEG data will not be presented here. A trial started with a short tone. Subsequently, a visual cue (“EE”) instructed the participants to vocalize the Dutch vowel /e/ for the duration of the cue (3s), while trying to match the pitch of the tone they just heard.

The participants were recorded, and the recorded voice signal was used to provide the participants with online auditory feedback. In half of the trials, participants received normal auditory feedback (control trials). In the other trials (perturbation trials), auditory feedback was normal at first, but starting between 500-1500ms after speech onset, the feedback's pitch was increased by 25 cents for a duration of 500 ms, before returning back to normal feedback for the remainder of the trial. Overall, participants received 99 perturbation trials and 99 control trials, randomly mixed in two blocks of 99 trials each.

4.2.3 Stimuli

The tone stimuli were pure tones with one of three frequencies. The pitch of the tones was individually tailored to the participants at 4, 8 and 11 semitones above the average pitch of five practice vocalizations.

The auditory feedback shifts were implemented using Audapter software (Cai, Boucek, Ghosh, Guenther, & Perkell, 2008). In brief, the software performs an online linear prediction coding analysis followed by a dynamic programming-based formant-tracking algorithm. To alter the formant frequencies, infinite-impulse-response filters are constructed and applied on the input waveform on the fly. The formant-shifted sounds are played back to the speaker with a latency of 10-20ms.

All recordings were made on one channel using a Sennheiser ME64 cardioid microphone, which was set up in the MEG magnetically shielded room and connected through an audio mixer to a dedicated soundcard Motu MicroBook II outside the room. Auditory feedback was delivered through the same soundcard which was connected to CTF audio air tubes. Stimulus presentation was controlled by a Windows computer running Audapter and MathWorks Matlab.

4.2.4 Analysis

For every trial of the speaking task, the pitch of participants' vocalization was determined using the autocorrelation method implemented in Praat (Boersma & Weenink, 2013). Subsequently, the pitch contours were exported to MATLAB for further processing.

Pitch contours were epoched from 500ms before perturbation onset to 1000ms after perturbation onset. For the control trials, in which there was no perturbation onset, random time points were chosen, while making sure the distribution of these time points across trials was equal to the distribution of perturbation onsets within the same participant. The data was detrended and converted from Hertz to the Cents scale using the following formula:

$$F0\ [cents] = 1200 \cdot log_2(F/F_{baseline})$$

Here, F is the original pitch frequency in Hertz, while $F_{baseline}$ is the average pitch frequency in Hertz across a baseline window (-200ms to 0ms before perturbation onset). Subsequently, trials that contained artifacts were removed from analysis. Artifacts were detected by visual inspection, looking for sharp discontinuities in the pitch contour, or the absence of a pitch contour.

Pitch contours in control and perturbation trials were compared using a cluster-based permutation test (Maris & Oostenveld, 2007). Samples for which the contrast Perturbation - Control exceeded an uncorrected α level of .05 were temporally clustered. Cluster-level statistics were calculated by summing the t-statistics. Next, a permutation distribution of statistics was calculated by randomly exchanging trials between the conditions, and calculating the maximal positive and negative cluster-level statistics for every permutation (total of 1,000 permutations). The observed cluster-level statistic was tested against the permutation distribution.

In contrast to most previous studies, each trial was classified as having an opposing or a following response. For this classification procedure, two different methods were used. In the first, the point was determined within the time window 60-400ms after perturbation onset at which the average of the rectified pitch contours was maximal (point of maximal deviation). Then a linear regression was performed on the single-trial data between 60ms after perturbation onset and the point of maximal deviation. If the slope of the linear fit was positive, the response was classified as following, if it was negative, the response was classified as opposing. For the distribution of the slopes of the linear fits, see the supplementary materials. No threshold was applied for the slope to be significantly different from 0, although additional analyses leaving out the trials with slope near 0 yielded similar results (see supplementary materials).

The second classification method was based on the Castellan change-point test (Siegel & Castellan, 1988). This change-point test yields the statistic K:

$$K=|2W_j - j(N+1)| \qquad\qquad j=1,2,\dots,N-1$$

Here, W_j is the cumulative sum of ranks at sample number j, and N is the total number of samples. We calculated K for every trial over the time window 0-300ms. The point where K is maximal is the change point. If $2W_j$ - j(N+1) at that point is positive, the trial was classified as opposing, if it was negative, the trial was classified as following.

If these two classification methods did not yield the same classification for a particular trial, classification was determined through visual inspection of the pitch contour (this occurred in on average 23.0% of a participant's data, range: 7.2- 43.3%). If there was no clear response, the trial was classified as having no response (on average, 7.9% (range: 2-16.3%) of trials were classified as having no response). This classification procedure was performed on the perturbation as well as on the control trials.

In order to look at how participants' responses depended on the state of their voice motor system at the moment where perturbation kicks in, for each trial type the slope and the average F0 value over the 100ms before perturbation onset were determined.

Another way to identify differences between opposing and following trials is to compare the magnitude and the latency of the responses. The peak response was identified for each trial type in each participant by subtracting the average pitch contour for the control trials from the average contour of the perturbation trials. The response latency was then quantified as the point in time between 50ms and 500ms at which the difference was largest.

4.3 RESULTS

Overall, participants compensated for the pitch increase in the perturbation trials by lowering their pitch (Figure 4.1a). The pitch contour in the perturbation trials differed from the control trials (p = 0.002). This difference was mainly driven by a component lasting from 144ms to 765ms after perturbation onset.

To investigate whether participants sometimes followed the feedback shift instead of opposing it, we classified each perturbation trial as either a following or an opposing trial (or neither). The same trial classification was performed on the control trials. The distribution of opposing and following trials (Figure 4.2) shows a clear effect of perturbation: in the control trials, the proportion of trials classified as opposing trials is about 50%, reflecting random fluctuations of the pitch contour, whereas in the perturbation trials, the proportion of opposing trials is larger (t(38) = 8.16, p < 0.001, CI = [0.14 0.23], Cohen's d = 1.96), ranging from just under 50% to over 90%. Clearly, participants followed the feedback perturbation in a non-trivial number of trials (10-50%).

As expected, the pitch contour in opposing trials differed from the following trials (Figure 4.1b, p = 0.002). This was mainly driven by a component (opposing < following) from 108ms to 812ms after perturbation onset, but also by a smaller difference in the opposite direction (following < opposing), from 91ms before perturbation onset until 77ms after perturbation onset. This suggests that even before perturbation onset, the pitch contours in these trials already differ.

The pitch contour in the opposing perturbation trials differed from the pitch contour in the opposing control trials (Figure 4.3a, p = 0.002). This was driven by a component (Figure 4.3c, 213ms-712ms) where pitch was lower in the perturbation trials and a later component (from 791ms) where

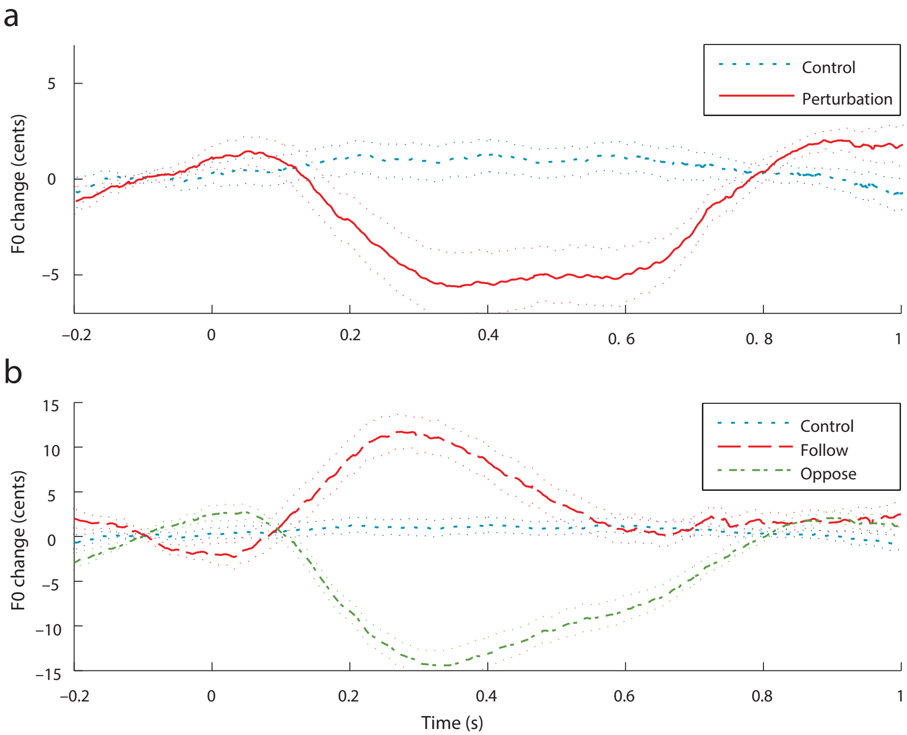

Fig. 4.1. Grand averages comparing change in F0 in perturbation and control trials, time-locked to perturbation onset (a). In the perturbation trials, perturbation starts at 0s and lasts until 0.5s. In (b), grand average for the perturbation trials split up in following and opposing trials. Thin dotted lines represent the 95% confidence intervals.

pitch in the perturbation trials was higher compared to the control trials. A similar pattern was found for the following trials (Figure 4.3b/d, p = 0.002), where the perturbation trials had lower pitch than the control trials from 338ms until 723ms after perturbation onset. So even for the perturbation trials classified here as following trials, the pitch was lower compared to the similarly classified control trials. This means that the pitch increase may not entirely be indicative of a following response, but may also (or instead) reflect an ongoing F0 fluctuation with an additional smaller opposing response.

The small early difference between following perturbation trials and opposing perturbation trials (Figure 4.1b) suggested a difference before

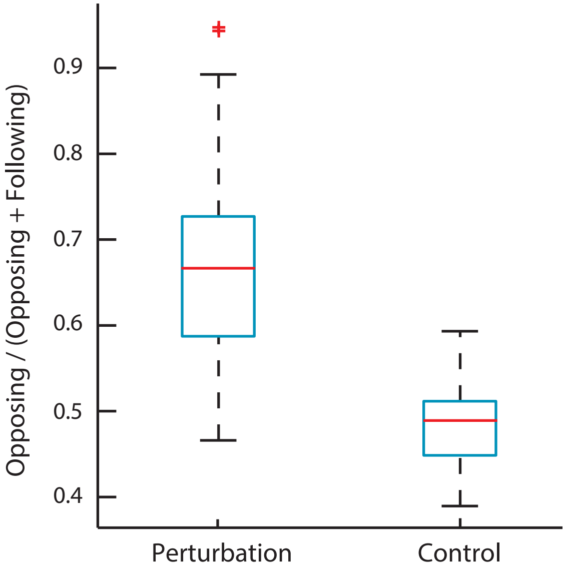

Fig. 4.2. Proportion of trials classified as opposing in perturbation and control conditions.

perturbation onset. The results of a Wilcoxon signed rank tests show that both the pitch slope (z = -4.24, p < 0.001, r = 0.48) and average value (z = -5.25, p < 0.001, r = 0.59) over the 100ms time window before perturbation onset differed between following and opposing trials (Figure 4.4). This effect was also found continuously across the data, as well as for the trials within each response type (see supplementary materials).

With respect to response peaks, following responses peaked on average earlier (t(38) = 3.66, p < 0.001, CI = [0.02 0.08], Cohen's d = 0.74) and were smaller (t(38) = 17.11, p < 0.001, CI =[24.28 30.80], Cohen's d = 3.91) than the opposing responses. An earlier and/or smaller response in following trials can be explained as a result of detection that the following response, on top of the feedback pitch shift, produces a pitch even further from the intended target.

4.4 DISCUSSION

The current study investigated speakers' responses to unexpected shifts in sensory feedback. An altered auditory feedback paradigm was used to investigate whether the response was dependent on the state of the speech

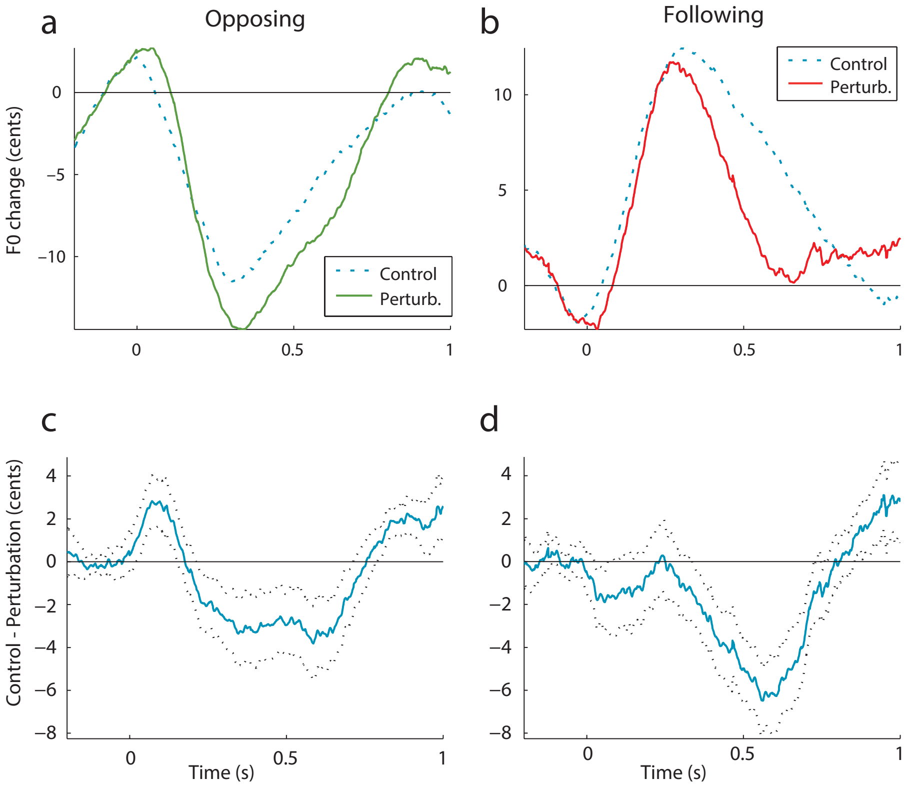

Fig. 4.3. Comparison of F0 change between perturbation and control condition for similarly classified trials. In (a), F0 change for opposing perturbation trials and control trials classified as opposing. In (b), F0 change for following perturbation trials and control trials classified as following. In (c) and (d), difference waves corresponding to the comparisons in (a) and (b); dotted black lines represent 95% confidence intervals.

production system at shift onset. Overall, participants compensated for the pitch-shifted feedback by opposing the direction of the pitch shift. This result is in line with previous research, and consistent with models that hypothesize that an internal forward model compares an incoming auditory signal with the predicted auditory feedback (Wolpert & Ghahramani, 2000). Interestingly, however, closer analysis revealed that all participants also followed the feedback shift on some trials. The proportion of following trials showed a lot of variability across subjects, ranging from about 10% to over 50%.

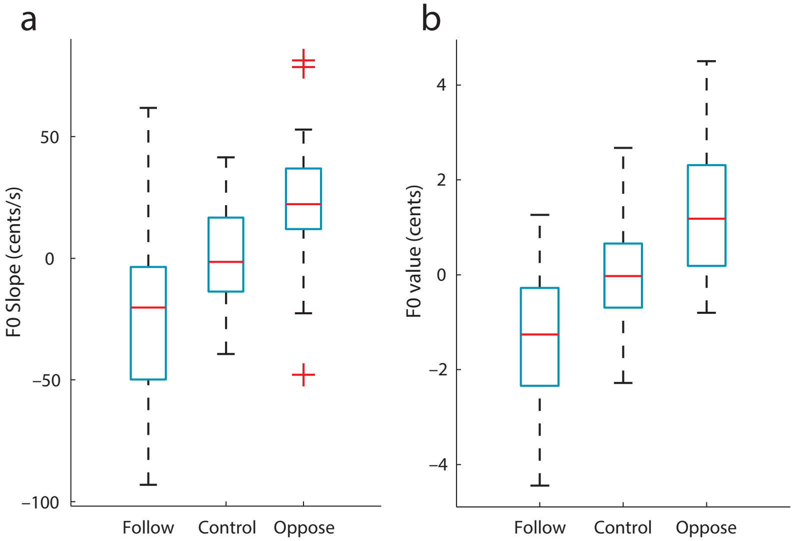

Fig. 4.4. F0 slope (a) and average F0 value (b) over a 100ms time window before perturbation onset as a function of trial classification.

Following responses are not in line with many models on sensory feedback processing for motor control (Houde & Nagarajan, 2011; Wolpert & Ghahramani, 2000). These models hypothesize that the goal of the motor system is to minimize the discrepancy between the predicted sensory representation and the sensory feedback. However, when participants follow the direction of feedback perturbations in their vocal output, the discrepancy between prediction and sensory input increases.

These models could account for following responses in two ways. One possibility is that participants may have difficulty determining the direction of the pitch shift. This difficulty may be due to the small magnitude of the pitch perturbation in this study; some listeners may be able to detect the pitch change without being able to correctly identify the direction of the pitch change (Neuhoff, Knight & Wayand, 2002). A misidentification of the direction may lead to following the pitch shift accidentally. However, as following responses are in fact less common with smaller feedback perturbations (Burnett et al., 1998), and given that Behroozmand et al. (2012) showed that following was more common when the pitch change direction was predictable, it is unlikely that a misidentification of the direction of the pitch perturbation is the sole cause of following responses.

Another possible account for following responses is that in some cases the auditory input is considered by the speaker to be externally-driven rather than self-generated. Such an explanation is in line with the model proposed by Hain et al. (2000), which proposes two feedback modes: one for tracking an external referent (hence following pitch changes), and another to compensate for internal disturbances (hence opposing pitch changes). The feedback mode is determined by distinguishing between self-generated and externally-generated auditory signals. Both modes could be active simultaneously, as for example during choir singing, which requires singers to compensate for disturbances in their own vocal output, while following changes in their fellow singers' output. In the current study, however, it is unclear why sometimes the feedback shift was considered to be self-generated (leading to opposing responses) and sometimes to originate from an external source (leading to following responses). In addition, the fact that no participant reported being aware of the perturbations suggest there were no clear changes in the perceived source of the auditory feedback.

If the present findings turn out to be robust, current models should be revised, because they cannot adequately account for following responses. Pitch is known to show fluctuations around a target pitch level (Akagi et al., 1998). These fluctuations could be driven by continuous feedback monitoring and (over)compensation. The current results indicate three interactions between ongoing fluctuations and the perturbation response.

First, the direction of the response was related to the slope and/or average value of the F0 contour before perturbation onset. This suggests that the response is dependent on the current state of the system. Specifically, when the participants' pitch was decreasing or was lower than average, the response tended to follow the pitch shift (increase in pitch), and vice versa. This suggests that the system initially reacts to the pitch perturbation by doing the opposite of what it's currently doing. When pitch is decreasing or lower than average, the system reacts by increasing pitch, and vice versa. This mechanism would be useful in a natural context. If a pitch mismatch is detected in the feedback signal, it is likely that ongoing compensatory articulations are going in the wrong direction. Thus simply changing pitch in the opposite direction would be a good strategy.

Second, comparing the perturbation trials to the similarly classified control trials, both ‘following’ trials as well as opposing trials showed an opposing trend, suggesting that for so-called following trials, an opposing response may be riding on top of on-going pitch fluctuations.

Third, the results showed that the response peak for opposing trials was larger and occurred later compared to following trials. After an initial strategy of simply changing pitch in the opposite direction to how it was changing before the pitch shift, a following response would be detected as increasing the prediction error even more, leading to a quicker readjustment, and thus an overall smaller response with an earlier peak.

Together, these findings show evidence of a dynamic interplay between the state of the motor system and incoming sensory feedback. This view is broadly in line with a dynamic systems approach to cognitive processing (Gelder, 1998). More generally, this study indicates that looking beyond the average response can lead to a more complete view on the nature of feedback processing in speech production and motor control. It also leads to the prediction that the direction of sensorimotor adaptation in domains outside speech production will also be conditional on the state of the motor system at the time of the perturbation.

APPENDIX

Response Slope

In order to classify every single trial as having either an opposing or a following response, we characterized the response slope, that is the slope of the pitch contour over the time window starting at 60ms after perturbation onset until the point of maximal deviation (the maximum of the average of rectified pitch contours). Figure 4.5 shows the distribution of these slopes across participants. The mean of the distribution was -31.70 (sd = 116.25). As described in more detail in the main text, given that the perturbation was always positive (+25 cents), a positive response slope was considered indicative of a following response while a negative slope was associated with an opposing response.

Continuous effect

For the analyses described in the main text, no threshold was applied to the response slopes before classification. In other words, only the sign of the slope was used for classification, not how much it deviated from 0. The main result of the paper, describing the dependency of the response type on the pre-perturbation pitch slope, is repeated here after having removed the trials with small slopes (i.e., close to 0) from the analysis.

Figure 4.4 shows that the slope (and average value) of the pitch contour right before perturbation onset in following trials is lower compared to opposing trials, suggesting the state of the pitch system may be a factor influencing the type of response to a pitch perturbation. In order to confirm these results, here we report the effect continuously in Figure 4.6.

Overall, it seems a higher (more positive) response slope is associated with a smaller (more negative) pre-perturbation slope in the pitch contour, and vice versa (Spearman's *rho* = -0.19, $p < 0.001$). When the data was split by response type, the same effect was found for both opposing trials (*rho* = -0.21, $p < 0.001$) and following trials (*rho* = -0.09, $p = 0.002$). Note that the effect was stronger for the opposing trials. This in line with available models as a lower than average pre-perturbation slope would be considered in and of itself already a small error (Niziolek, Nagarajan, & Houde, 2013), and given that the pitch perturbation was positive, a lower-than-average pitch value would only increase the resulting prediction error, thus leading to a stronger (opposing) response.

To make sure the effect is not driven by trials with a very small or no response (i.e., a response slope close to 0, see above), the analyses were repeated on the 60% most extreme data points (i.e., the 30% with the most positive response slope and the 30% with the most negative response slope). The results are shown in Figure 4.7. The results are similar to the analysis on

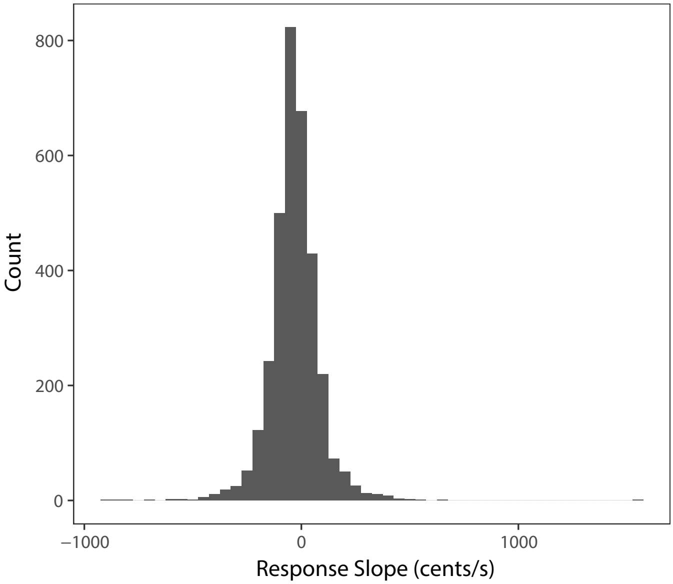

Fig. 4.5. Histogram of the response slope (pitch contour slope from 60ms after perturbation onset until the point of maximal deviation) across participants. A negative slope indicates decreasing pitch and thus an opposing trial (given a perturbation of +25 cents), a positive slope indicates a following trial.

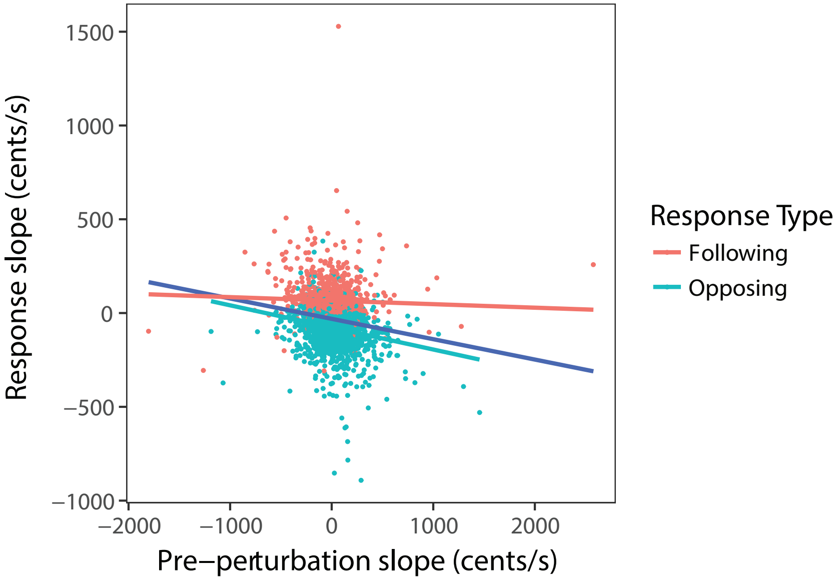

Fig. 4.6. Scatter plot of the response pitch contour slope as a function of pre-perturbation pitch slope, across single trials. Colors indicate the response type classification of the trial. The dark blue linear fit is the fit to the complete dataset.

all the data. Again, an overall association was found between response slope and pre-perturbation slope (*rho* = -0.21, *p* < 0.001), as well as for just the opposing trials (*rho* = -0.17, *p* < 0.001) and a weaker one for the following trials (*rho* = -0.12, *p* = 0.006).

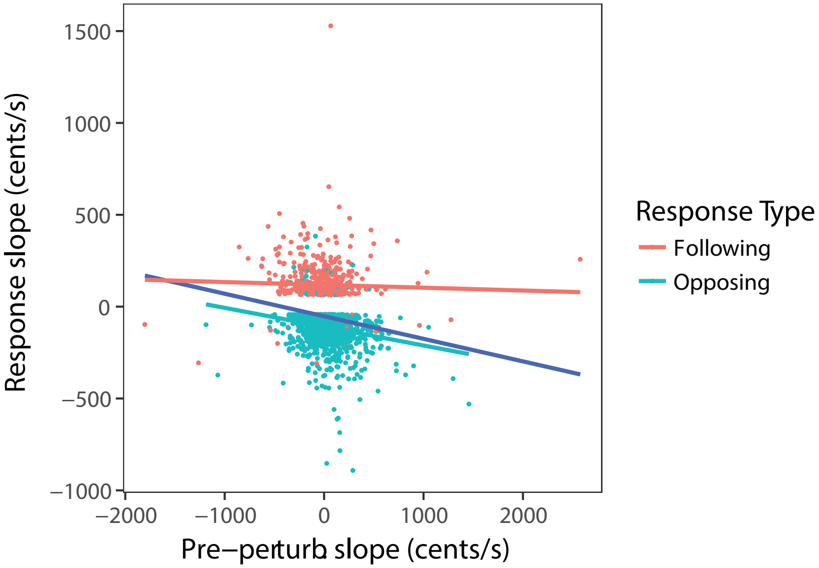

Fig. 4.7. Scatter plot of the response pitch contour slope as a function of pre-perturbation pitch slope, across single trials, excluding 40% of the dataset around a response slope of 0 (indicating no response to the perturbation). Colors indicate the response type classification of the trial. The dark blue linear fit is the fit to the complete dataset.

REFERENCES

Akagi, M., Iwaki, M., & Minakawa, T. (1998). Fundamental frequency in continuous vowel utterance and its perception. *Proc. ISCLP98*, pp. 1519-1522.

Akagi, M., & Kitakaze, H. (2000). Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours. *Proc. ICSLP2000*, pp. 458-461.

Behroozmand, R., Korzyukov, O., Sattler, L., & Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: evidence for different mechanisms of voice pitch control. *The Journal of the Acoustical Society of America, 132*(4), 2468–77. doi:10.1121/1.4746984

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*(6), 3153–3161. doi:10.1121/1.423073

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In *Proceedings of the 8th Intl. Seminar on Speech Production* (pp. 65–68). Strasbourg, France.

Ford, J., Gray, M., Faustman, W., Heink, T., & Mathalon, D. (2005). Reduced gamma-band coherence to distorted feedback during speech when what you say is not what you hear, *International Journal of Psychophysiology, 57*(2), pp. 143-150.

Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Brain and behavioural sciences, 21*(5), 615-628.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301. doi:10.1016/j.bandl.2005.06.001

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research, 130*(2), 133–141. doi:10.1007/s002219900237

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135–145. doi:Doi 10.1038/Nrn2158

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5(28). doi:10.3389/fnhum.2011.00082

Larson, C. R., Sun, J., & Hain, T. C. (2007). Effects of simultaneous perturbations of voice pitch and loudness feedback on voice F0 and amplitude control. *The Journal of the Acoustical Society of America, 121*(5), 2862. doi:10.1121/1.2715657

Liu, H., & Larson, C. R. (2007). Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex. *The Journal of the Acoustical Society of America, 122*(6), 3671–7. doi:10.1121/1.2800254

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience methods, 164*(1), 177-190. Doi:10.1016/j.neumeth.2007.03.024

Neuhoff, J. G., Knight, R., & Wayand, J. (2002). Pitch change, sonification, and musical expertise: which way is up? In *Proceedings of the 2002 International Conference on Auditory Display*. Kyoto, Japan.

Niziolek, C., Nagarajan, S., & Houde, J. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience, 33*(41), 16110-16116.

Siegel, S., & Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd ed., p. 399). New York.

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nat Neurosci, 3*(Suppl), 1212–1217.

5

AUDITORY FEEDBACK IN THE CEREBRAL CORTEX

ABSTRACT

Speaking is a complex motor skill, requiring integration of sensory and motor-related information. Current theory hypothesizes a complex interplay between motor and auditory processes during speech production, which among others involves the online comparison of the speech output with an internally generated forward model. Over the last couple of decades, perturbed auditory feedback has been used to study this. The current study examines the neural correlates of processing and responding to altered auditory feedback. Participants vocalized the vowel /e/ and listened to auditory feedback that was temporarily pitch-shifted, while brain signals were recorded with magneto-encephalography (MEG). Afterwards, participants listened to recordings of the same auditory feedback they were exposed to in the first half of the experiment, now without vocalizing. Participants were not aware of any auditory perturbation. We observed strong auditory cortical responses to both perturbation onset and offset during speech production, but not during listening, suggesting that auditory cortex was especially sensitive to small pitch shifts during speech production. In addition, auditory feedback perturbation resulted in power increases in the θ and lower β bands. It is suggested that both θ and β band activity can be related to how auditory and motor processing. are integrated. Overall, these results are in line with current models of speech production, in suggesting a comparison between a forward model's prediction and auditory feedback in auditory cortices, which subsequently interact with motor areas to generate a motor response. Furthermore, the results extend current knowledge about the neural correlates of auditory feedback processing, by suggesting that θ and β power increases support auditory-motor interaction, motor error detection and/or sensory prediction processing.

5.1 INTRODUCTION

Speaking is a remarkably complex motor skill. We speak with a rate of about fifteen speech sounds per second, each of which require accurate coordination of more than 100 different muscles. We make use of this skill day in day out, throughout our lives, usually without conscious awareness of the complexity of the task. If at all, attention is mostly paid to wording, while articulation follows effortlessly. In order to perform this motor task almost without errors, a good quality control system is needed. Recent developments in speech motor control have shown that integration of sensorimotor information, including auditory feedback (i.e. the sound of our own voice), is key in this respect. The current study investigates the neural underpinnings of sensorimotor integration during speech production.

The role of auditory feedback in speech production has been investigated by providing speakers with online manipulated feedback (Houde & Jordan, 1998; Jones & Munhall, 2000). For example, speakers could be hearing their own speech in real time at a higher pitch or with a lower first formant. It turns out that speakers usually compensate for these unexpected manipulations by changing their speech in the opposite direction, even when told to ignore the altered feedback (Keough, Hawco, & Jones, 2013). This suggests that speakers automatically monitor their auditory feedback during speech production. In order to explain such a fast feedback monitoring mechanism, modeling work in speech production has drawn from principles in motor control more generally (Wolpert, Ghahramani, & Jordan, 1995; Wolpert & Ghahramani, 2000). These models hypothesize the use of internally generated forward models (Houde & Nagarajan, 2011; Tourville & Guenther, 2011). Specifically, whenever an articulatory motor program is generated, which is to be executed by the motor system, an efference copy is sent to the auditory system. This efference copy can be used by the forward model, which models the sensory (auditory) consequences of the articulation. This sensory prediction can then be compared with the observed sensory consequences, and if necessary generate a prediction error that could signal the need for behavioral adaptation.

Researchers have also used the altered auditory feedback paradigm to study the neural correlates of feedback processing (Behroozmand & Larson, 2011; Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Zarate & Zatorre, 2005). Several functional magnetic resonance imaging studies have shown that feedback processing is supported by an extended functional neural network including auditory and motor-related areas bilaterally (Behroozmand, Shebek, et al., 2015; Zarate, Wood, & Zatorre, 2010; Zheng et al., 2013; Zheng, Munhall, & Johnsrude, 2010). Electrophysiological research has started to explore the temporal dynamics of feedback processing. Studies using electroencephalography (EEG) have shown that altered feedback leads to a brain response as early as 100ms after perturbation onset (Behroozmand, Karvelis, Liu, & Larson, 2009; Behroozmand, Liu, & Larson, 2011; Hawco, Jones, Ferretti, & Keough, 2009). In particular, the event-related components N1 and P2 were related to various aspects of feedback processing. These early auditory components have been found to be associated with the direction as well as the magnitude of the perturbation (Behroozmand & Larson, 2011; Liu, Meshman, Behroozmand, & Larson, 2011; Scheerer, Behich, Liu, & Jones, 2013), with the acoustical complexity of the feedback (Behroozmand, Korzyukov, & Larson, 2011), with age (Scheerer, Liu, & Jones, 2013) and with musical experience or skill (Behroozmand, Ibrahim, Korzyukov, Robin, & Larson, 2014; Korzyukov, Karvelis, Behroozmand, & Larson, 2012). The early latency of these effects suggests that auditory processing is already interacting with motor control at an early processing stage.

However, only a small number of studies on feedback perturbations have looked beyond evoked responses. This may be surprising, as recent dynamic approaches to cognition have linked cortical oscillations to predictive processing (Engel, Fries, & Singer, 2001) and sensorimotor integration more generally (Caplan et al., 2003), as well as to speech production specifically (Cruikshank, Singhal, Hueppelsheuser, & Caplan, 2012; Gehrig, Wibral, Arnold, & Kell, 2012; Jenson et al., 2014). In a recent study, power in both θ and δ bands was found to increase after a feedback pitch perturbation (Behroozmand, Ibrahim, Korzyukov, Robin, & Larson, 2015). Specifically, the authors report a θ band increase that overlapped with the behavioral response, and a later δ increase starting roughly 1s after the perturbation. A magnetoencephalography (MEG) study reported increased gamma power in response to feedback pitch shifts in left sensory and right premotor, parietal and temporal regions (Kort, Cuesta, Houde, & Nagarajan, 2016). These results indicate that power increases over motor and sensory areas may reflect sensorimotor speech processing.

In order to expand knowledge of feedback processing during speech production, the present study used MEG to investigate further the neural correlates of pitch perturbation processing and of subsequent automatic responses to such perturbations. A very small perturbation magnitude (25 cents) was used to make sure the participants did not consciously detect the perturbation. Although Behroozmand et al. (2015) suggested that the θ band activity increase during feedback perturbation they found is related to enhanced automatic pitch processing, their study used a pitch perturbation that may very well be consciously detected by the listener (100 cents). In fact, they explicitly made their participants aware of the perturbation by asking them to not produce any voluntary vocal responses to the onset of pitch stimuli. If their findings indeed reflect automatic pitch processing, we should find similar neural responses with smaller perturbations that are not consciously detected. In addition, while the majority of past research on pitch perturbations has indeed applied a fairly large perturbation, some studies have shown that attentional factors can influence participant's responses (Hu et al., 2015; Korzyukov, Sattler, Behroozmand, & Larson, 2012; Liu et al., 2015). This raises the question how conscious perception of and attention to altered feedback may affect feedback-based speech production. In fact, some studies have shown that perturbation responses are automatic (Burnett, Freedland, Larson, & Hain, 1998) and may even occur without conscious perception (Hafke, 2008). Hafke (2008) argued for a dissociation between a (conscious) perception stream and an action stream, where the latter includes auditory feedback processing for action purposes. It is however not entirely clear how these two streams interact. In fact, Hain et al. (2000) show that speakers can voluntarily change or modify at least part of their responses to altered feedback. The current study used a small perturbation in order to avoid attentional effects or conflation of the perception and action streams in auditory processing.

Interestingly, several authors have also reported that sometimes participants do not alter their speech output in response to the feedback perturbation by opposing the direction of the perturbation, but by following it. For example, if they hear themselves at a higher than expected pitch, they would increase their pitch. To date, it is unclear why participants would sometimes follow feedback perturbations instead of opposing them. The previous chapter describes a behavioral analysis of the current experiment that targeted the distinction between opposing and following responses specifically. The current chapter focuses on the neural correlates of auditory feedback processing. An additional analysis, in line with the previous chapter, will investigate whether neural activity can help explain the difference between following and opposing behavioral responses.

5.2 MATERIALS AND METHODS

5.2.1 Subjects

Thirty-nine healthy volunteers (age: M = 22, range = 18-34; 27 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local ethics board committee (CMO region Arnhem / Nijmegen). All participants had normal hearing, were native speakers of Dutch and had no history of speech and/or language pathology.

5.2.2 Paradigm

An experimental session consisted of two tasks, a speaking and a listening task, always performed in the same order (speaking, then listening), while brain activity was measured using MEG.

In the speaking task, participants performed a tone-matching task. A trial started with the presentation of a short tone. Subsequently, a visual cue (“EE”, in Dutch pronounced as /e/) instructed the participants to start vocalizing /e/, while trying to match the pitch of the tone they just heard. The visual cue disappeared after 3s, cueing the participant to stop vocalizing. During speech production / vocalization, the participant's voice was recorded using a microphone, positioned about 1,5m from the participant to avoid any artifacts in the MEG signal. The recorded signal was used to provide the participants with online auditory feedback. In half of the trials, participants received normal auditory feedback throughout the trial (henceforth control trials). In the other half of the trials (perturbation trials), auditory feedback was normal at first, but, starting between 500-1500ms after speech onset (randomly jittered), the feedback's pitch was increased by 25 cents for a duration of 500ms, before returning back to normal feedback for the remainder of the trial. Overall, participants received 99 perturbation trials and 99 control trials, randomly mixed in two blocks of 99 trials each. After the speaking task, participants did a passive listening task, in which the participants were shown the same visual cues as in the production task, but were instructed not to speak. Instead, they listened to recordings of the very same feedback they were given in the speaking task.

Finally, after the experiment, participants filled out a short debriefing questionnaire, which asked whether they noticed any feedback manipulations and if so, what kind of manipulations.

5.2.3 Materials

The tone stimuli were pure tones at one of three pitch frequencies. The pitch of the tones was individually tailored to the participants at 4, 8 and 11 semitones above their conversational pitch. This was done by having participants produce the vowel /e/ five times (they were not yet aware the experiment would involve pitch), and the average pitch was considered their conversational pitch.

The auditory feedback shifts were implemented using Audapter software (Cai, Boucek, Ghosh, Guenther, & Perkell, 2008; Tourville, Cai, & Guenther, 2013). In brief, the software performs a near-real-time autocorrelation analysis to track the pitch. In order to shift the pitch, the short-time Fourier spectra are stretched and interpolated along the frequency axis. The pitch-shifted sounds are played back to the speaker through earphones or headphones with a latency of 10-20ms.

All voice recordings were made on one channel using a Sennheiser ME64 cardioid microphone, which was set up in the MEG magnetically shielded room and connected through an in-house-built audio mixer to a dedicated soundcard Motu MicroBook II outside the room, which was connected to a Windows computer. Auditory feedback was delivered through the same soundcard which was connected to CTF (VSM/CTF systems, Port Coquitlam, Canada) audio air tubes. Stimulus presentation and sound recording times were controlled by the same Windows computer running Audapter and MathWorks Matlab (MathWorks, Version 8 Release 5, Natick, MA).

5.2.4 MEG Acquisition

We used an MEG system (VSM/CTF systems, Port Coquitlam, Canada) with 275 axial gradiometers. Three localization coils, fixed to anatomical landmarks (nasion, left and right preauricular points), were used to determine head position relative to the gradiometers. Head position was monitored online by the experimenter and if necessary corrected between the experimental blocks. All data were low-pass filtered by an anti-aliasing filter (300 Hz cut-off), digitized at 1200 Hz and stored for offline analysis. Participants were seated upright, with the head rested against the back of the helmet and touching the top of the helmet. A small cushion was used to fix the head's position so as to minimize free head movement. The participant's head movement and position was monitored in realtime and, if necessary, adjusted between blocks (Stolk, Todorovic, Schoffelen, & Oostenveld, 2013). was A headband was used to cover the audio air tubes and the participants' ears, minimizing the effect of air-conducted auditory feedback.

5.2.5 MRI Acquisition

In order to reconstruct the sources of the sensor-level MEG results, T1-weighted anatomical MRI scans were acquired for 34 out of the 39 subjects. Scans were acquired using Siemens 1.5T Avanto scanner for 24 participants, a Siemens 3T Prisma scanner for 6 participants, and a Siemens 3T Skyra scanner for 4 participants, depending on scanner availability.

5.2.6 Analysis

5.2.6.1 Behavioral

For every trial of the speaking task, the pitch of participants' vocalization was determined using the autocorrelation method implemented in Praat (Boersma & Weenink, 2013). Subsequently, the pitch contours of all trials were exported to MATLAB (The MathWorks Inc., Natick, MA, 2012) for further processing.

Pitch contours were epoched from 500ms before perturbation onset to 1000ms after perturbation onset. For the control trials, in which there was no perturbation onset, random time points were chosen, while making sure the distribution of these time points across trials was equal to the distribution of perturbation onsets within the same subject. The data was de-trended and converted from Hertz to the Cents scale using the following formula:

$$F0\ [cents]=1200 \cdot \log_2(F/F_{baseline})$$

Here, F is the original pitch frequency in Hertz, while $F_{baseline}$ is the average pitch frequency in Hertz across a baseline window (-200ms to 0ms before perturbation onset). Subsequently, trials that contained artifacts were removed from analysis. Artifacts were detected by visual inspection, looking for sharp discontinuities in the pitch contour, or the absence of a pitch contour.

5.2.6.2 MEG preprocessing

All analyses were performed in MATLAB (The MathWorks Inc., Natick, MA, 2012), using custom scripts and the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011).

First, data was epoched from 1s before speech onset (or audio onset in the listening task) to 6s after speech onset. Bad channels were removed, and the data was de-meaned and visually inspected for artifacts. Segments containing artifacts were removed. Subsequently, an ICA algorithm (Hyvärinen, 1999) was performed to identify eye movement and heartbeat artifacts. The time course and topography of the ICA components that showed highest coherence with EOG and ECG channels were inspected and components showing artifacts were removed from the data. On average, about 6 components were removed for each subject (ranging from 3 to 9 components). Subsequently, the remaining components were projected back to sensor space.

5.2.6.3 ERF

MEG data was time-locked to perturbation onset, or, for the control trials, to a randomly chosen time point (see behavioral analysis). Every trial was filtered using a zero-phase forward windowed sinc FIR filter with a Hamming window and a 1-40Hz passband. Subsequently, the data was cut into time windows from -1s to 2s after perturbation onset, de-trended, averaged per condition and per participant and transformed to synthetic planar gradients (Bastiaansen & Knosche, 2000).

An additional analysis examined the neural correlates of the distinction between following and opposing responses to the altered auditory feedback. Trials were classified as having either an opposing or a following response (see chapter 4). As the distribution of trials was very uneven between the opposing and following classes, the following procedure was used to enable statistical comparison between the neural responses for opposing and following trials. For every participant, the minimum number of trials in a response class was determined (most often this was in the following-response class). Five random subsets of that number of trials were selected from the other response class. The event-related field (ERF) response for that response class was calculated by averaging the ERF across the five trial subsets. This way, each ERF was calculated by averaging, within each participant, over the same number of trials in both response classes.

5.2.6.4 TFR

For the time-frequency analyses, the data (-1s to 2s after perturbation onset) was de-meaned and transformed to the frequency domain using a sliding 500ms Hanning tapered window, sliding in steps of 50ms from -500ms to 1500ms after perturbation onset. The frequency band of interest ranged from 2Hz up to 30Hz (in steps of 2Hz). Before transformation, the input was zero-padded to 4s.

5.2.6.5 Statistical inference

For statistical inference, a non-parametric permutation test was performed with a clustering method to correct for multiple comparisons (Maris & Oostenveld, 2007). This was done for the data 0-1s after perturbation onset to determine whether there was a difference between the perturbation and the control condition within the speaking task. Samples for which the contrast Perturbation - Control exceeded an uncorrected α level of .05 were spatiotemporally clustered. Cluster-level statistics were calculated by summing the t-statistics. Next, a permutation distribution of statistics was calculated by randomly exchanging data between the conditions, and calculating the maximal positive and negative cluster-level statistics for every permutation (for a total of 1,000 permutations). The observed cluster-level statistic was tested against the permutation distribution.

In order to compare the ERF results of the speaking and the listening tasks across conditions, the average activity was calculated for every subject in both tasks and both conditions on a 100ms time window centered at the point of maximal t-value (averaged across channels) for the contrast perturbation-control in the speaking task at group level. This was done both for the maximal t value after perturbation onset (170ms) and after perturbation offset (627ms). The resulting average activity values were entered in 2-way repeated measures ANOVAs (one for the averages after perturbation onset, and one for perturbation offset), with factors Task (speak vs. listen) and Condition (perturbation vs. control). Post-hoc t-tests were carried out to compare the perturbation and control trials within the listening task.

5.2.6.6 MRI processing

In order to estimate source-level activity, we co-registered the anatomical MRI to the MEG sensors. This was achieved by identifying in the MRI the anatomical locations that were used to place the head localization coils during the MEG measurement (left/right ear canal, and nasion). Subsequently, the aligned image was used to create (1) a volume conduction model based on a single shell model of the inner surface of the skull, and (2) a description of the cortical surface, using Freesurfer 5.1 (Dale, Fischl, & Sereno, 1999). The individual cortical surfaces were surface-registered to a template and downsampled to 4002 nodes per hemisphere, using the Connectome Workbench software (http://www.humanconnectome.org/connectome/connectome-workbench.html).

5.2.6.7 Beamforming

The sensor-level results were projected onto the individual cortical surfaces using beamforming techniques. Data visualization was performed using the Connectome Workbench of the Human Connectome Project (http://www.humanconnectome.org/connectome/connectome-workbench.html). For the event-related data, we used a time-domain beamformer (LCMV). The data covariance was calculated across a time window ranging from -150ms to 800ms after perturbation onset across both (perturbation and control) conditions. Spatial filters were calculated, based on the forward solution, and a regularized inverse of the covariance matrix, averaged across conditions (the regularization parameter was set to 10% of the average sensor signal variance). Next, for each condition separately, the common spatial filter was used to estimate the source activity for three time windows of interest: perturbation onset (100-250ms), perturbation offset (550-700ms) and intermediate (300-400ms).

For the time-frequency results, a frequency-domain beamformer (DICS) was used. The data was de-meaned and the cross-spectral density was calculated over a 1.5s time window (-500ms to 1,000ms), across conditions, centered on 7Hz for the θ band (bandwidth 4-10Hz), and on 17Hz for the β band (bandwidth 14-20Hz), using dpss tapers. The resulting cross-spectral densities were combined with the forward solution to calculate frequency band-specific spatial filters (regularization parameter was at 10%). Next, condition-specific cross-spectral densities were calculated over the time window 0-500ms, and combined with the common spatial filters to obtain condition-specific source estimates.

5.3 RESULTS

5. 3.1 Behavioral responses

Overall, participants compensated for the pitch increase in the perturbation trials by lowering their pitch (Figure 5.1). A cluster-based permutation test revealed that participants' pitch contour in the perturbation trials was different from the control trials (*p* = 0.002). This difference was mainly driven by a cluster lasting from 144ms to 765ms after perturbation onset. For a more extended analysis of the behavioral results in terms of opposing and following responses, see chapter 4. Results from the debriefing questionnaire revealed that none of the participants was aware of any pitch perturbations in the auditory feedback.

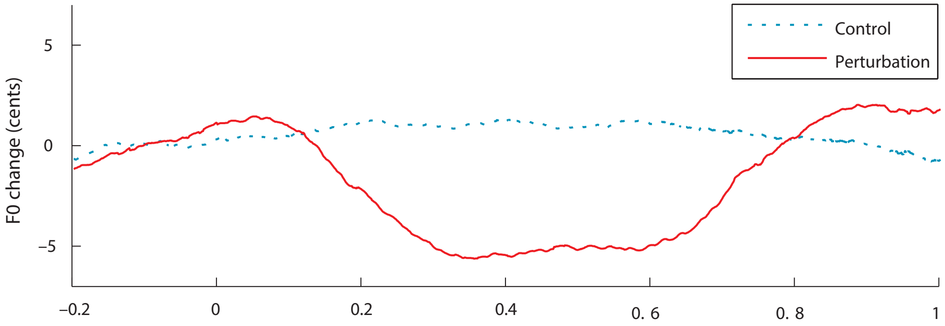

Fig. 5.1. Average pitch contour in perturbation and control trials. For the perturbation trials, the perturbation started at 0s and lasted until 0.5s.

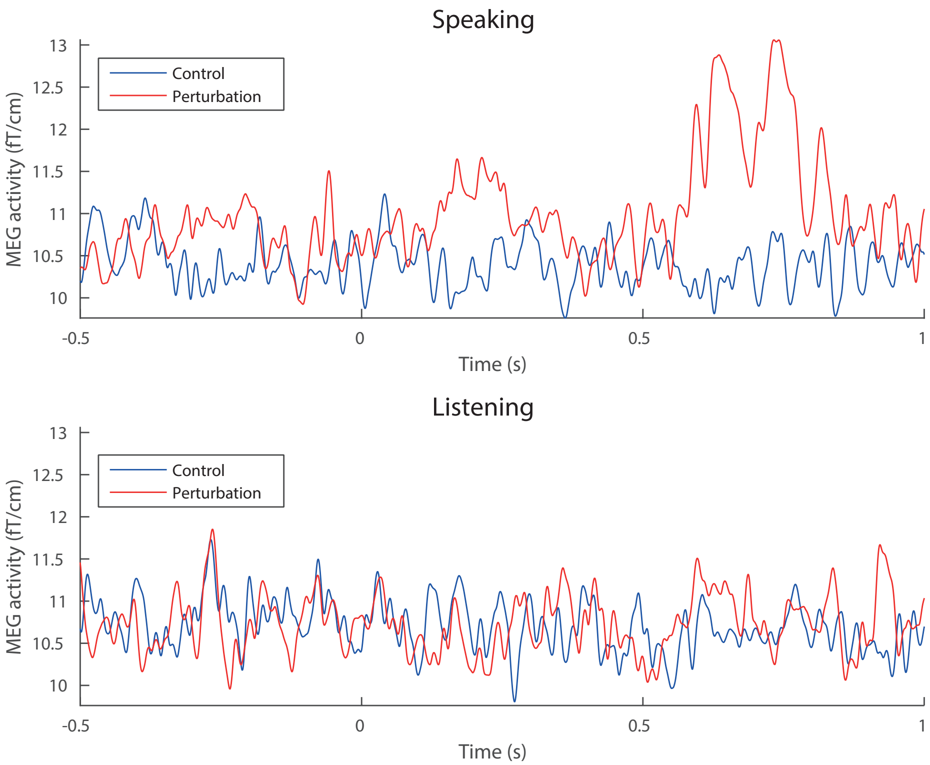

Fig. 5.2. Event-related field of perturbation (red) and control (blue) trials, averaged across all channels, for the speaking task (top graph) and the listening task (bottom graph). For the perturbation trials, the perturbation started at 0s and lasted until 0.5s.

5.3.2 ERF

The main analyses collapsed over both behavioral response types (following and opposing). Overall, the event-related fields show a response to both perturbation onset and offset in the speaking task, but not in the listening task (Figure 5.2). A cluster-based permutation test on the speaking task data within 1s after perturbation onset revealed a significant difference between perturbation and control trials (p < 0.001). This difference was mainly driven by an increase in activity in the perturbation condition after both perturbation onset (from about 85ms after speech onset to about 250ms) and perturbation offset (550ms-850ms).

The topography plots for the speaking task (Figure 5.3) show a mainly right-lateralized pattern in both time windows. A smaller left-lateralized cluster of increased activity in perturbation vs. control trials was found in

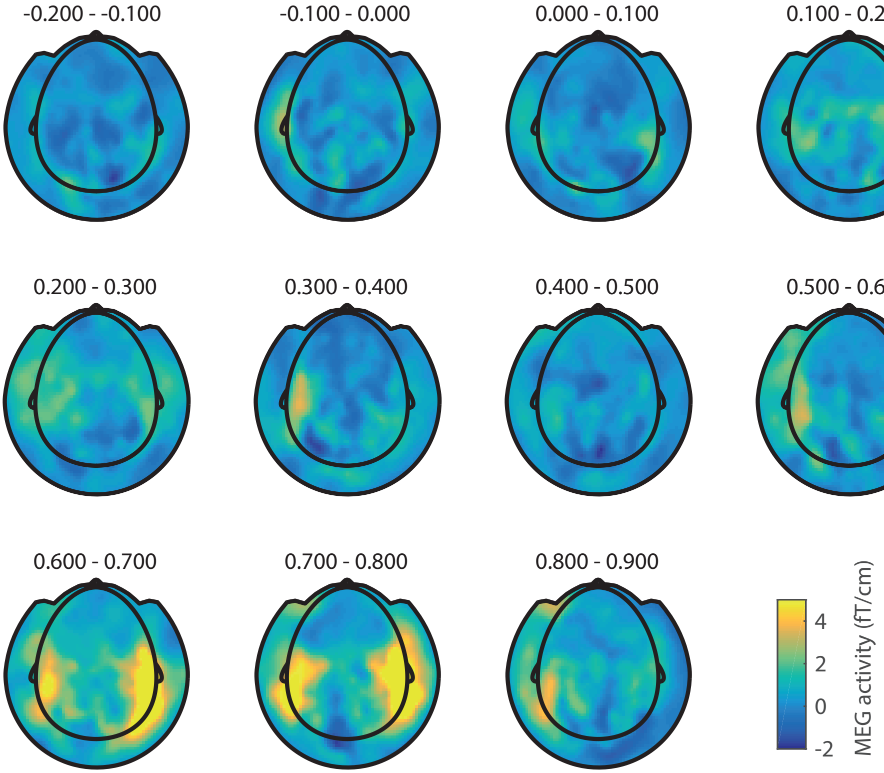

Fig. 5.3. Topography plots of the difference (perturbation – control) in the speaking task per 100ms, from -200ms to -100ms (top left) to 800ms-900ms (bottom right). For the perturbation trials, the perturbation started at 0s and lasted until 0.5s.

a later time window (300-400ms). There was no clear similar cluster after offset that emerged from the cluster analysis, although note that the main cluster after perturbation offset lasted relatively long, until about 850ms after perturbation onset, that is 350ms after perturbation offset. Figure 5.3 suggests that activity in this last part of the offset-related cluster (800-900ms) was also left-lateralized. The topography plots for the listening task (Figure 5.4) show that there is little difference, if anything, between the perturbation and control conditions.

A comparison of MEG activity in speaking and listening tasks across conditions (Figure 5.5) showed an interaction between task and condition in both the onset-related time window ($F(1, 35) = 17.362$, $p < 0.001$) and the offset-related time window ($F(1, 35) = 15.449$, $p < 0.001$). Post-hoc t-tests within the listening task showed that neither the difference between perturbation and control conditions in the onset-related time window ($t(1, 35) = 2.01$, *n.s.*, uncorrected), nor the difference in the offset-related time window ($t(1, 35) = -1.68$, *n.s.*, uncorrected) led to a significant overall change in MEG activity.

An LCMV beamformer was used to project the sensor-level activity of the speaking task in three windows of interest (onset: 100-250ms; offset: 550-700ms; and a third time window: 300-400ms) onto the cortical surface. The results are depicted in Figure 5.6. Both perturbation onset and perturbation offset-related activity increases were localized to superior temporal and inferior frontal areas, lateralized to the right hemisphere. Activity over the

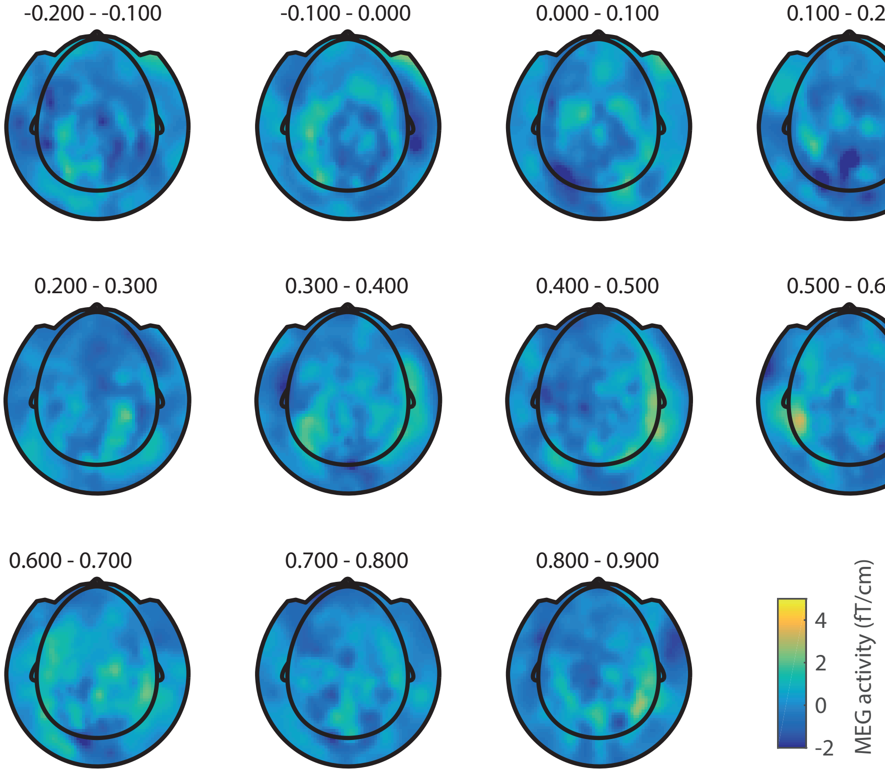

Fig. 5.4. Topography plots of the difference (perturbation – control) in the listening task per 100ms, from -200ms to -100ms (top left) to 800ms-900ms (bottom right). For the perturbation trials, the perturbation started at 0s and lasted until 0.5s.

300-400ms time window showed increased activity in areas around the central sulcus in the left hemisphere.

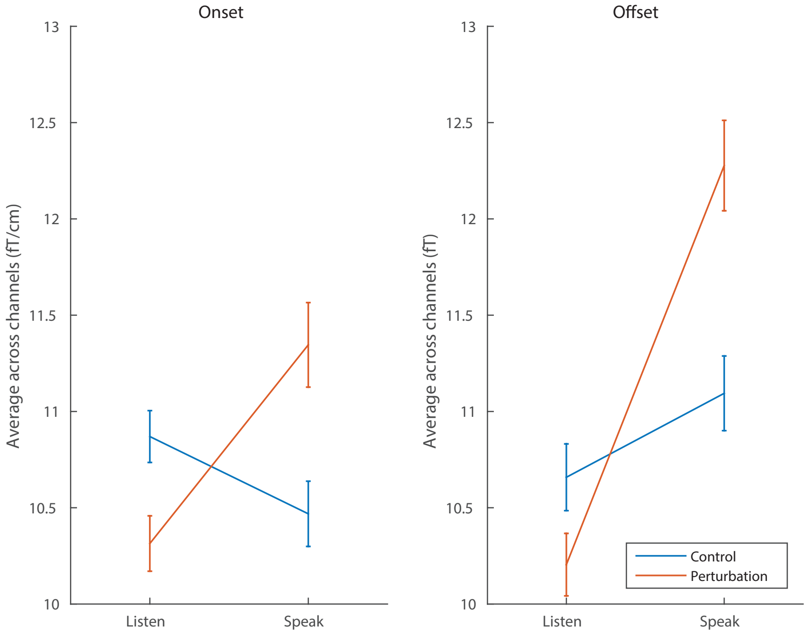

Fig. 5.5. Average activity across channels in a 100ms time window after perturbation onset (left, centered at 170ms) and after perturbation offset (right, centered at 627ms). Error bars indicate standard errors. There was a significant interaction between task and condition in both time windows.

5.3.3 TFR

A time-frequency analysis of the data time-locked to perturbation onset shows event-related power changes across the low frequency range. Figure 5.7 shows the contrast (uncorrected t-values) of these power changes between perturbation and control trials. A cluster-based permutation test revealed that there was a significant power increase in the perturbation condition, relative to the control condition (p = 0.041), which was mainly driven by increased power in the θ (4-8Hz) and a lower β (12-16Hz) band.

Topography (Figure 5.8) and source-level (Figure 5.9) plots indicate involvement of sensorimotor areas. The results of the DICS beamformer (Figure 5.9) suggest that θ band activity was associated mostly with areas around inferior primary motor and somatosensory cortical areas (parts of Brodmann areas 1, 2, 3, 4 and 6), whereas the lower β band power increase

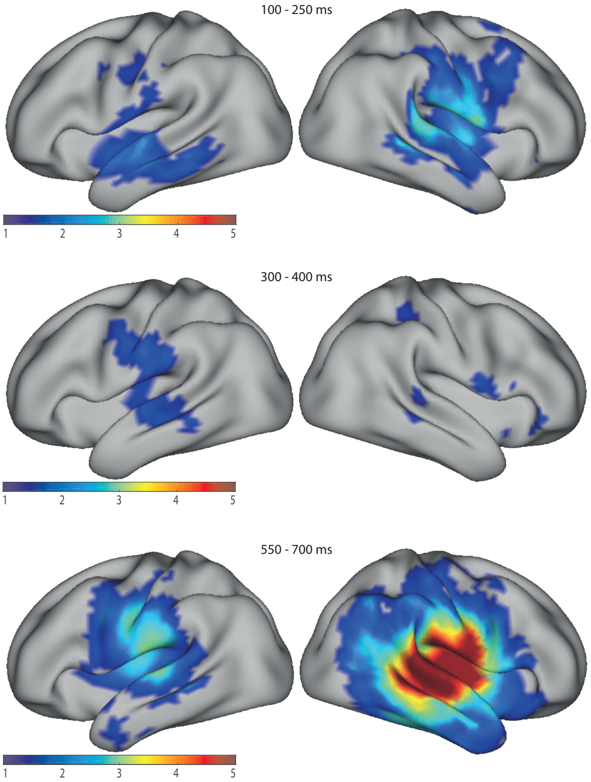

Fig. 5.6. Source-level event-related fields over the time windows 100-250ms (top), 300-400ms (middle) and 550-700ms (bottom). Right hemispheres on the right and left hemispheres on the left. The colors indicate the contrast (perturbation - control) / baseline and are thresholded.

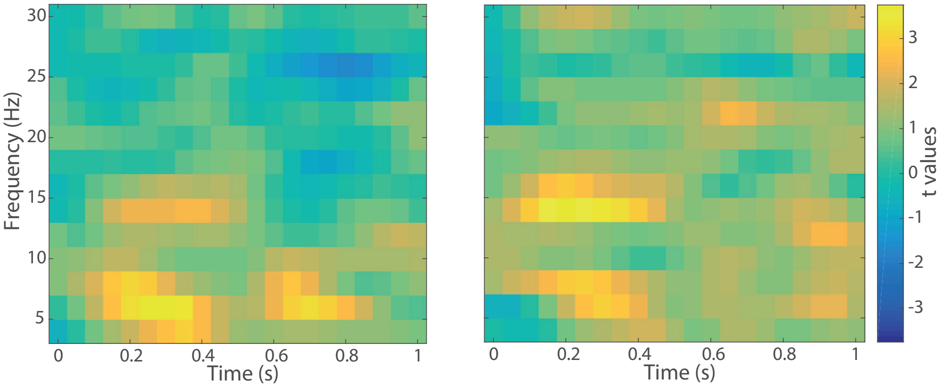

Fig. 5.7. Average t values indicating power changes in the perturbation condition, relative to the control condition, across the lower frequencies. Data was time-locked to perturbation onset. The left graph shows the power changes averaged across 10 channels (see marked channels in Figure 5.8) that were especially sensitive to the θ power difference (4-10 Hz, 0-0.5s), the right graph does the same for the β window (14-20Hz, 0-0.5s).

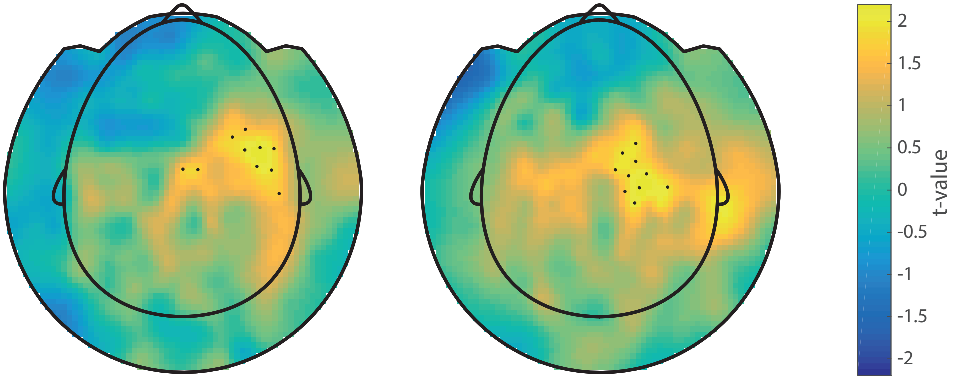

Fig. 5.8. Topography plots of θ band (4-10Hz, left) and β band (14-20Hz, right) power increase in perturbation trials compared to control trials. Colors indicate average t-values over 0-500ms after perturbation onset. The channels selected for the average spectrograms in Figure 5.7 are marked.

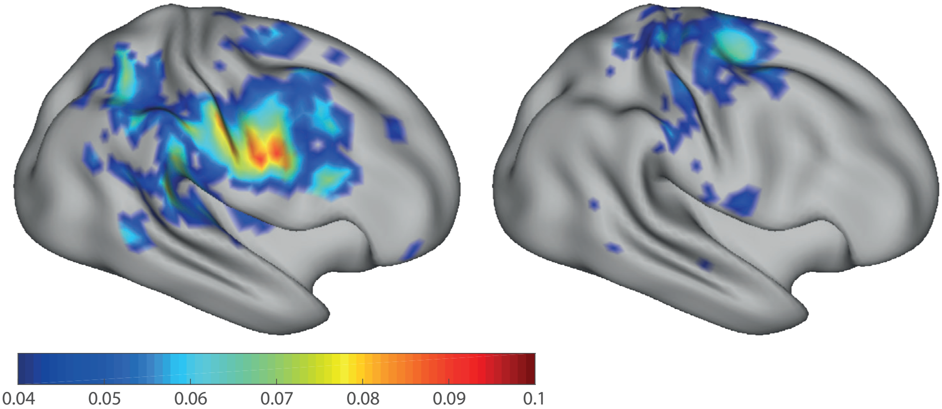

Fig. 5.9. Source-level projections of θ band (left, 4-10Hz, 0-500ms) and β band (right, 14-20Hz, 0-500ms) perturbation-related power increases on the right hemisphere.

was projected onto more superior motor areas (parts of Brodmann areas 4 and 6).

5.3.4 Neural correlates of behavioral response type

As a secondary analysis, trials were classified as following (when the participant's behavioral response followed the direction of the feedback pitch perturbation) or as opposing (when the participant behaviorally opposed the pitch perturbation). Figure 5.10 shows the event-related field responses corresponding to opposing and following trials. A cluster-based permutation test revealed the ERF for following and opposing trials significantly differed from each other (p = 0.02). From the figure, it can be observed that this difference was mainly driven by a difference in the activity over central channels from 100 to 250ms after perturbation onset. A second, smaller, cluster that showed up between roughly 550ms-650ms was hard to interpret in the current paradigm given its posterior location. Also, it would

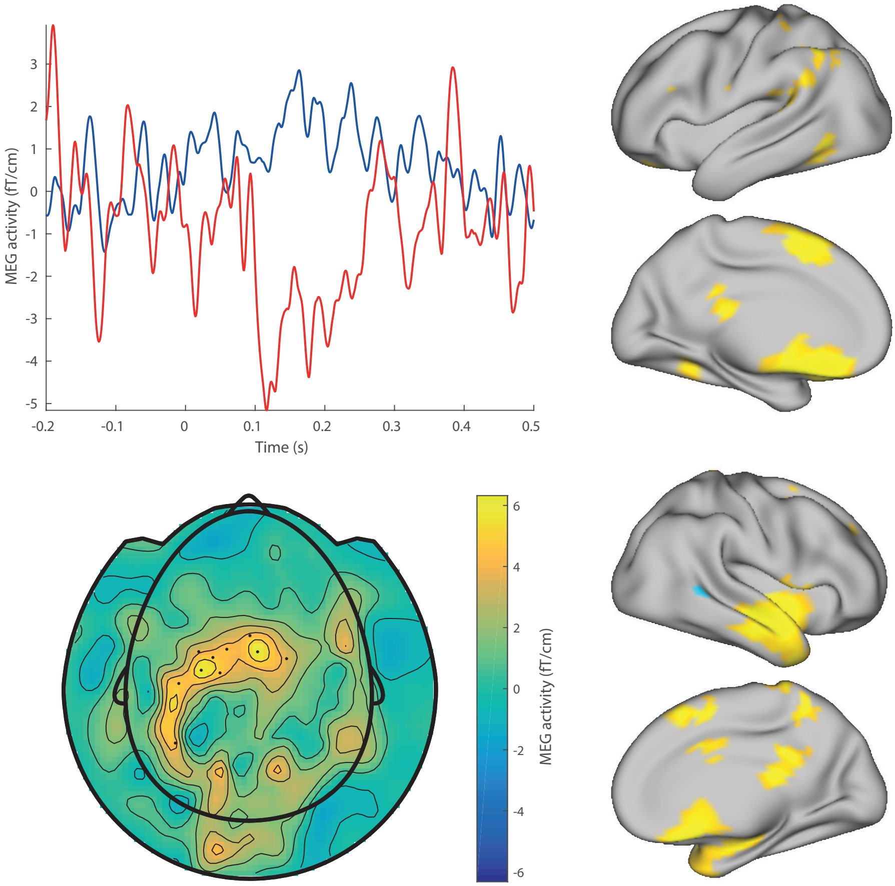

Fig. 5.10. Top left: event-related field responses to opposing (blue) and following (red) trials, averaged across the channels highlighted in the topography plot. Perturbation onset is at 0ms. Bottom left: topography plot of the condition difference opposing – following, average over the time window 100-250ms. Highlighted channels are the channels used for the top left plot. Right: source plots (top to bottom: left lateral view, left medial, right lateral, right medial) of the results of a LCMV beamformer. Colors indicate the difference (opposing – following)/(common_baseline) and are thresholded at values -4 and 4.

be hard to interpret neural activity around 550ms-650ms that is linked to an earlier behavioral response. The topography of our main cluster (100ms-250ms) suggests the activity difference originates from motor-related areas. The results of a beamforming analysis (shown on the right side of Figure 5.10) suggests the motor-related area involved may be the supplementary motor area (SMA). Other areas, mainly the bilateral ventromedial prefrontal cortices (vmPFC) and areas in the right middle temporal lobe, show up in the beamforming analysis, though these are not as clear from the results of the sensor-level topography plot. In addition, given that MEG is less sensitive to activity in deeper brain areas (compared to EEG or fMRI), especially the activity in vmPFC should be interpreted with caution. The activity in the right temporal lobe may be related to sensorimotor processing, although it is located more anterior compared to our main findings (Figure 5.6). Note that the areas in this additional analysis in general do not overlap with our main analysis, which collapsed across following and opposing trials. This indicates that our main findings are valid for both response types (following and opposing), and that activity in these areas, including SMA and anterior temporal areas, are especially sensitive to the response types, without showing an overall (response type-independent) effect.

5.4 DISCUSSION

In the current study, the neural correlates of perceiving and responding to an unexpected feedback pitch shift were investigated. While none of the participants were consciously aware of the pitch shifts, they did respond to the perturbation, both behaviorally and at the neural level. This suggests that both types of response reflect an unconscious auditory processing stream (Hafke, 2008). At the neural level, results showed a strong time-locked response in auditory areas to both perturbation onset and offset, as well as power increases in both the θ and the lower β band during the perturbation. These power increases were localized to frontal motor-related areas.

Behaviorally, participants responded to the perturbation by compensating for about 20% of the pitch shift on average. This finding is consistent with the vast literature on altered auditory feedback, and supports cognitive models hypothesizing that sensory feedback is continuously monitored to update and maintain adequate motor commands, both within and outside the domain of speech production (Houde & Nagarajan, 2011; Tourville & Guenther, 2011; Wolpert & Ghahramani, 2000).

At the neural level, event-related fields showed a response to both perturbation onset and offset, as well as a smaller left-lateralized response at a longer latency after perturbation onset (and a similar one after perturbation offset, though not analyzed). Source reconstructions suggest that the onset- and offset-related responses were generated around bilateral auditory cortical areas, with a stronger effect in the right hemisphere. The right-lateralized effect is in line with the well-established view that the right hemisphere is dominant in pitch processing (Johnsrude, Penhune, & Zatorre, 2000; Zatorre, Evans, Meyer, & Gjedde, 1992).

In line with previous studies of pitch perturbations in auditory feedback (Behroozmand & Larson, 2011; Liu et al., 2011), the results presented in the current study and their associated source reconstructions suggest the onset- and offset-related responses include, but may not be limited to, detection of the pitch shift and/or prediction-feedback discrepancy by auditory and surrounding sensorimotor areas. Interestingly, these responses are not solely due to auditory detection, as participants did not show the same results in the listening task. This suggests that auditory cortical areas are especially tuned to these pitch shifts during speech production. Note that previous studies of pitch change detection or detection of auditory mismatches in general find similar patterns. EEG studies of auditory mismatch detection have found the so-called mismatch negativity (MMN), a negativity response to an auditory oddball that peaks about 150-250ms after stimulus onset (Näätänen, Gaillard, & Mäntysalo, 1978; Näätänen, Paavilainen, Rinne, & Alho, 2007). This timing corresponds well to what we find here, although multiple components may overlap. Functional magnetic resonance imaging has suggested that these MMNs are generated by a number of different sources, including superior temporal and surrounding areas (Molholm, Martinez, Ritter, Javitt, & Foxe, 2005), which corresponds well to the source reconstructions of the onset- and offset-related responses reported here.

The ERF results reported here show, surprisingly, a stronger response to the offset of perturbation, compared to the onset of perturbation. In the light of an internal forward model, the ERF peaks reflect a comparison between the observed speech output and the predicted outcome as generated by the forward model. As the speaker adapted his/her speech output in order to minimize the prediction error due to the perturbation, the offset of the perturbation effectively introduces a new (reversed) perturbation. This new perturbation then again would trigger the detection of a pitch change at the neural level, although the difference in magnitude with the onset-related response is unexpected. If we consider the perturbation onsets and offsets as two consecutive pitch shifts, three aspects may be important in the distinction between them: (1) the order (the mere fact that the offset pitch shift is the second, and therefore is following another pitch shift), (2) regularity (the presence and timing of the onset pitch shift is unpredictable, but if there is an onset pitch shift, there will always be an offset pitch shift 500ms later), and (3) time delay from speech onset. We may relate this to the phenomenon of repetition enhancement, where the neural response to a repeated stimulus is enhanced (as opposed to the perhaps more widely known repetition suppression). Although the factors that determine whether repetition leads to suppression or enhancement are not so clearcut (Segaert, Weber, de Lange, Petersson, & Hagoort, 2013), attention has been shown to be able to boost neural activity (Corbetta & Shulman, 2002; Nakamura, Dehaene, Jobert, Le Bihan, & Kouider, 2007). Segaert et al. (2013) also suggest that there are good theoretical reasons to expect selective attention could lead to enhancement effects. Note that although our participants were not aware (even after the experiment) of any pitch manipulations, the pitch matching task explicitly drew their attention to their pitch, and a first, unexpected, pitch shift might have unconsciously triggered more attention to the pitch tracking task, and hence to the pitch of the auditory feedback. Although usually it has been found that a regular second stimulus (like our second shift, the perturbation offset) leads to neural suppression, maybe the absence of conscious awareness of the pitch shifts does not trigger repetition suppression. In addition, the pitch shifts in this experiment are of course not repetitions in the strict sense. In fact, the second pitch shift has the opposite direction, which perhaps leads to an enhancement rather than neural suppression. Finally, the second pitch shift simply occurs later, that is there is a longer delay between speech onset and perturbation offset compared to perturbation onset. As the pitch contour is more variable close to speech onset, the perturbation onset shift may be less salient to the (unconscious) speech processing machinery, compared to the perturbation offset.

The smaller ERF peak at 300-400ms after perturbation onset was localized to left (pre)motor and somatosensory areas. This suggests that this response may be related to the motor response, that is, to the behavioral compensatory response. There was no clear similar left-lateralized cluster after perturbation offset, but the early offset-related cluster may have included this response, especially given the left-lateralized topography in 800-900ms in Figure 5.3. Although most research has focused on a right-lateralized network involved in pitch-shifted feedback processing, some fMRI studies have reported activity in similar left-hemisphere areas. In one study, participants vocalized /a/ for 5s while their feedback pitch could be randomly shifted up or down from the original pitch two or three times in a trial (Toyomura et al., 2007). Participants were explicitly instructed to compensate for these feedback shifts. Increased activity in the left premotor area was reported in the perturbation condition (compared to a control condition). Behroozmand, Shebek, et al. (2015) used a similar experimental paradigm and found (amongst other regions) increased activity in the left superior temporal gyrus during 600 cent pitch shifts. Interestingly, the BOLD response in the bilateral superior temporal gyri and the left precentral gyrus was positively correlated with the magnitude of the vocal response. This suggests that responses in these areas are related to the behavioral response. Although most electrophysiological studies have used much shorter pitch shifts (mostly 200 ms) and do not report responses at these latencies, Kort, Nagarajan & Houde (2014) have reported some speaking-specific responses at longer latencies as well. In an MEG study, they report increased activity in response to a feedback pitch shift in left superior temporal gyrus, left ventral supramarginal gyrus, as well as in left and right premotor cortices. As these responses occur after the onset of the behavioral response, they cannot be related to preparing the vocal response. Kort et al. (2014) suggest that this might reflect the efference copy of the corrected motor plan being fed back to the auditory cortex. Note that more general, speech production models like DIVA (Tourville & Guenther, 2011) and the state feedback control model (Houde & Nagarajan, 2011) also include these areas, although without lateralization. For example, DIVA posits articulator velocity and position maps in the bilateral ventral motor cortices, while the state feedback control model (Houde & Nagarajan, 2011) posits the premotor cortices as an intermediary, supporting prediction and correction processes running between motor and sensory areas. Therefore, we suggest the left-lateralized response found in this study may be related to the implementation of a behavioral/motor response to the perturbation.

Overall, the results from the event-related field analyses show right-lateralized responses in auditory cortices and surrounding areas to both pitch perturbation onset and offset, as well as a left-lateralized response in motor-related areas around 300ms after perturbation onset. A similar response could be seen around 300ms after perturbation offset (800-900ms in Figure 5.3). These results suggest an interconnected sensory-motor network that supports auditory-motor integration, including auditory and motor-related areas in both hemispheres.

The present study adds to the literature by investigating not only the event-related neural effects, but also by looking at both evoked and induced effects in a time-frequency analysis. The results showed evidence of increased θ and β band power during and/or after feedback pitch perturbation. The θ band power increase is in line with the suggestion of Behroozmand et al. (2015), that increased θ band power reflects "mechanisms by which humans incorporate auditory feedback to control their voice pitch" (p. 10). The source-level projection of the θ band increase in the current study suggested a source in inferior motor areas and/or the posterior temporal areas. This is in line with the hypothesis that auditory and motor-related areas are jointly involved in feedback-based vocal pitch adjustments. In addition, the current study extends the previous finding by Behroozmand et al. (2015) by using small feedback perturbations (so listeners were not aware of the perturbation), as well as by projecting the sensor-level data onto the cortical surface, and therefore showing the θ band power is primarily localized to lower motor and somatosensory cortical areas.

Furthermore, a power increase in the lower β band was found. This frequency band showed a power increase that did not extend beyond the perturbation offset, in contrast to the θ band result. To the best of our knowledge, a β band increase has not been described previously in an altered auditory feedback paradigm. One reason may be that many previous studies have used much shorter pitch perturbations (100-200ms), and thus are less equipped to demonstrate reliable β modulation during the perturbation. However, β band neural activity has been studied in relation to both motor and auditory processing more generally. Specifically, Koelewijn, van Schie, Bekkering et al. (2008) have suggested that β oscillations in the motor cortex are modulated by the correctness of observed actions. Participants in their study were asked to observe actions (button presses) that could be correct or incorrect. They found that while β power was suppressed during the action, β power showed a stronger rebound for incorrect actions compared to correct ones. The authors suggest that the increased β rebound may reflect active inhibition of the motor system after the detection of an erroneous action. Relating that to the present study, one might speculate that increased β power could reflect the detection of erroneous pitch production (thus an incorrect speech action). The source projection also supports an interpretation as involvement in motor activity and/or planning, with the main source located in more superior (pre)motor cortical areas.

Modulation of β power has also been relatively well documented in the literature on auditory beat processing (Fujioka, Trainor, Large, & Ross, 2009; Iversen, Repp, & Patel, 2009). Specifically, Iversen et al. suggest that β band oscillations play a role in motor-auditory interactions, and especially in motor cognition effects on auditory perception. Auditory-motor interaction is a crucial aspect of the current paradigm, as an unexpected auditory event (pitch perturbation) led to motor adjustments. The increase in β power could therefore be reflecting additional motor-auditory interactions as a result of the unexpected pitch shift. Interestingly, a recent EEG study found a change in induced β power to be related to an unpredicted pitch change in a rhythmic tone sequence (Chang, Bosnyak, & Trainor, 2016). They found a β increase after a tone that deviated in pitch from what was expected, and suggest β oscillations play a role in sensory prediction for both what will occur as well as when it will occur. In the present study as well, β oscillations could be related to the prediction of the sensory consequences of a vocalization action (i.e., the forward model), and thus a β increase may reflect the detection of a prediction error with regards to the sensory predictions. Note, however, that the source of the β increase in the present study was in motor areas, while most studies of auditory β have shown the β increase in auditory cortex or surrounding areas.

So two possible accounts of the role of β power in the current paradigm emerge: (1) error detection similar to error monitoring in the action literature or (2) auditory prediction as suggested by the literature on auditory beat processing. Future studies could try to disentangle these hypotheses. If increased β power is related to auditory prediction, it should be modulated by predictability of the pitch perturbation (see for example the consistency manipulation in chapter 3), while it should remain stable if it reflects only action error monitoring. Note that both accounts seem in contrast to the recently proposed hypothesis that β oscillations reflect the maintenance of the current cognitive set (Engel & Fries, 2010; Lewis & Bastiaansen, 2015; Lewis, Schoffelen, Schriefers, & Bastiaansen, 2016). Maintaining the current cognitive set seems at odds with a change in ongoing vocalization due to an unexpected change in auditory feedback. On the other hand, one could argue that increased β band activity in the current study reflects an increased need of maintaining the current articulatory goals (e.g., the target pitch) in the context of unexpected feedback. Future research should examine more closely to what extent β band activity under altered auditory feedback is at odds with the β maintenance hypothesis.

Finally, in line with the behavioral results reported in chapter 4, a secondary analysis investigated the neural correlates of the type of behavioral response to the pitch shift. The results showed an increased ERF response for the opposing responses (or a decrease for the following responses), shortly after perturbation onset. The locus of this effect seemed to include the supplementary motor area (SMA), amongst some other areas, such as the right middle temporal lobe. According to the DIVA model, SMA is involved in an initiation circuit, which ensures that articulations start at the right time and are timed correctly. In the current study, increased SMA activity during opposing responses may possibly signal the initiation of an opposing behavioral response. However, also in the following trials there was a behavioral response, though simply in the other direction. It is unclear why SMA activity distinguished between the two trial types. It is possible that following responses do not require initiation of a new, compensatory action, but instead reflect simple ongoing convergence to an external auditory stimulus, while opposing responses are generated by the initiation of a new articulatory action. Further research should try to clarify what the role of the increased SMA activity is with respect to following vs. opposing behavioral responses. The results in the previous chapter suggested the difference between following and opposing responses may be related to the state of the speech system at perturbation onset. An experiment where the perturbation would be conditional on the state of the speech production system could clarify whether SMA activity varies as a function of the system's state. Note that there were no other differences between following and opposing responses, suggesting that the main results reported in the current chapter are representative of both following and opposing trials.

The current study explored the neural underpinnings of auditory feedback processing during speech production. The study found that even without conscious awareness, speakers compensate for unexpected pitch shifts in auditory feedback. At the neural level, a strong short-latency response is found in auditory cortices reflecting the detection of the unexpected pitch. At a longer latency, neural activity associated with preparation or implementation of motor compensation was observed in left pre-motor areas. In addition, a power increase in both θ and β bands occurred as a response to the pitch perturbation. The θ power effect concurs with the literature and suggests the involvement of mechanisms which incorporate auditory feedback in voice control. We extend this literature by showing that the increased θ power is indeed related to automatic, unconscious, pitch processing, and by localizing it to motor-related cortical areas. To the best of our knowledge, this study is the first that shows an increase in β power in an altered auditory feedback paradigm. Increased β power may reflect motor error-monitoring or auditory prediction mechanisms. Overall, the results reported here are in line with current models of speech production, which posit that auditory feedback is constantly monitored during speech production. Even small unexpected errors are quickly detected by the auditory system and may lead to subsequent behavioral changes through increased auditory-motor interaction.

REFERENCES

Bastiaansen, M. C. M., & Knosche, T. R. (2000). Tangential derivative mapping of axial MEG applied to event-related desynchronization research. *Clinical Neurophysiology, 111*(7), 1300–1305. doi:Doi 10.1016/S1388-2457(00)00272-8

Behroozmand, R., Ibrahim, N., Korzyukov, O., Robin, D. A., & Larson, C. R. (2014). Left-hemisphere activation is associated with enhanced vocal pitch error detection in musicians with absolute pitch. *Brain and Cognition, 84*(1), 97–108. doi:10.1016/j.bandc.2013.11.007

Behroozmand, R., Ibrahim, N., Korzyukov, O., Robin, D. A., & Larson, C. R. (2015). Functional role of delta and theta band oscillations for auditory feedback processing during vocal pitch motor control. *Frontiers in Neuroscience, 9*, 109. doi:10.3389/fnins.2015.00109

Behroozmand, R., Karvelis, L., Liu, H., & Larson, C. R. (2009). Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. *Clinical Neurophysiology, 120*(7), 1303–1312. doi:http://dx.doi.org/10.1016/j.clinph.2009.04.022

Behroozmand, R., Korzyukov, O., & Larson, C. R. (2011). Effects of voice harmonic complexity on ERP responses to pitch-shifted auditory feedback. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 122*(12), 2408–17. doi:10.1016/j.clinph.2011.04.019

Behroozmand, R., & Larson, C. (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. *BMC Neuroscience, 12*(1), 54. Retrieved from http://www.biomedcentral.com/1471-2202/12/54

Behroozmand, R., Liu, H., & Larson, C. R. (2011). Time-dependent neural processing of auditory feedback during voice pitch error detection. *Journal of Cognitive Neuroscience, 23*(5), 1205–17. doi:10.1162/jocn.2010.21447

Behroozmand, R., Shebek, R., Hansen, D. R., Oya, H., Robin, D. A., Howard, M. A., & Greenlee, J. D. W. (2015). Sensory-motor networks involved in speech production and motor control: an fMRI study. *NeuroImage, 109*, 418–28. doi:10.1016/j.neuroimage.2015.01.040

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America, 103*(6), 3153–3161. doi:10.1121/1.423073

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In *Proceedings of the 8th Intl. Seminar on Speech Production* (pp. 65–68). Strasbourg, France.

Caplan, J. B., Madsen, J. R., Schulze-Bonhage, A., Aschenbrenner-Scheibe, R., Newman, E. L., & Kahana, M. J. (2003). Human theta oscillations related to sensorimotor integration and spatial learning. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 23*(11), 4726–4736. doi:23/11/4726 [pii]

Chang, A., Bosnyak, D. J., & Trainor, L. J. (2016). Unpredicted Pitch Modulates Beta Oscillatory Power during Rhythmic Entrainment to a Tone Sequence. *Frontiers in Psychology*, 7(MAR), 1–13. doi:10.3389/fpsyg.2016.00327

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences of the United States of America, 110*(7), 2653–2658. doi:DOI 10.1073/pnas.1216827110/-/DCSupplemental

Corbetta, M., & Shulman, G. L. (2002). Control of Goal-Directed and Stimulus-Driven Attention in the Brain. *Nature Reviews Neuroscience, 3*(3), 215–229. doi:10.1038/nrn755

Cruikshank, L. C., Singhal, A., Hueppelsheuser, M., & Caplan, J. B. (2012). Theta oscillations reflect a putative neural mechanism for human sensorimotor integration. *Journal of Neurophysiology, 107*(1), 65–77. doi:10.1152/jn.00893.2010

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis I. Segmentation and Surface Reconstruction. *NeuroImage, 9*, 179–194.

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology, 20*(2), 156–165. doi:10.1016/j.conb.2010.02.015

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience, 2*(10), 704–716. doi:10.1038/35094565

Fujioka, T., Trainor, L. J., Large, E. W., & Ross, B. (2009). Beta and gamma rhythms in human auditory cortex during musical beat processing. *Annals of the New York Academy of Sciences, 1169*, 89–92. doi:10.1111/j.1749-6632.2009.04779.x

Gehrig, J., Wibral, M., Arnold, C., & Kell, C. A. (2012). Setting up the speech production network: how oscillations contribute to lateralized information routing. *Frontiers in Psychology, 3*, 169. doi:10.3389/fpsyg.2012.00169

Hafke, H. Z. (2008). Nonconscious control of fundamental voice frequency. *The Journal of the Acoustical Society of America, 123*(1), 273–8. doi:10.1121/1.2817357

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research, 130*(2), 133–141. doi:10.1007/s002219900237

Hawco, C. S., Jones, J. A., Ferretti, T. R., & Keough, D. (2009). ERP correlates of online monitoring of auditory feedback during vocalization. *Psychophysiology, 46*(6), 1216–1225. doi:10.1111/j.1469-8986.2009.00875.x

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9469813

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*(28). doi:10.3389/fnhum.2011.00082

Hu, H., Liu, Y., Guo, Z., Li, W., Liu, P., Chen, S., & Liu, H. (2015). Attention modulates cortical processing of pitch feedback errors in voice control. *Scientific Reports, 5*, 7812. doi:10.1038/srep07812

Hyvärinen, A. (1999). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Trans. on Neural Networks, 10*(3), 626–634.

Iversen, J. R., Repp, B. H., & Patel, A. D. (2009). Top-down control of rhythm perception modulates early auditory responses. *Annals of the New York Academy of Sciences, 1169*, 58–73. doi:10.1111/j.1749-6632.2009.04579.x

Jenson, D., Bowers, A. L., Harkrider, A. W., Thornton, D., Cuellar, M., & Saltuklaroglu, T. (2014). Temporal dynamics of sensorimotor integration in speech perception and production: independent component analysis of EEG data. *Frontiers in Psychology, 5*, 656. doi:10.3389/fpsyg.2014.00656

Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000). Functional specificity in the right human auditory cortex for perceiving pitch direction. *Brain, 123*(1), 155–163. doi:10.1093/brain/123.1.155

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America, 108*(3), 1246. doi:10.1121/1.1288414

Keough, D., Hawco, C., & Jones, J. a. (2013). Auditory-motor adaptation to frequency-altered auditory feedback occurs when participants ignore feedback. *BMC Neuroscience, 14*, 25. doi:10.1186/1471-2202-14-25

Koelewijn, T., van Schie, H. T., Bekkering, H., Oostenveld, R., & Jensen, O. (2008). Motor-cortical beta oscillations are modulated by correctness of observed action. *NeuroImage, 40*(2), 767–775. doi:10.1016/j.neuroimage.2007.12.018

Kort, N. S., Cuesta, P., Houde, J. F., & Nagarajan, S. S. (2016). Bihemispheric network dynamics coordinating vocal feedback control. *Human Brain Mapping*. doi:10.1002/hbm.23114

Kort, N. S., Nagarajan, S. S., & Houde, J. F. (2014). A bilateral cortical network responds to pitch perturbations in speech feedback. *Neuroimage, 86*(0), 525–535. doi:http://dx.doi.org/10.1016/j.neuroimage.2013.09.042

Korzyukov, O., Karvelis, L., Behroozmand, R., & Larson, C. R. (2012). ERP correlates of auditory processing during automatic correction of unexpected perturbations in voice auditory feedback. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology, 83*(1), 71–8. doi:10.1016/j.ijpsycho.2011.10.006

Korzyukov, O., Sattler, L., Behroozmand, R., & Larson, C. R. (2012). Neuronal mechanisms of voice control are affected by implicit expectancy of externally triggered perturbations in auditory feedback. *PLoS ONE, 7*. doi:10.1371/journal.pone.0041216

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex, 68*, 155–168. doi:10.1016/j.cortex.2015.02.014

Lewis, A. G., Schoffelen, J.-M., Schriefers, H., & Bastiaansen, M. (2016). A Predictive Coding Perspective on Beta Oscillations during Sentence-Level Language Comprehension. *Frontiers in Human Neuroscience*, 10. doi:10.3389/fnhum.2016.00085

Liu, H., Meshman, M., Behroozmand, R., & Larson, C. R. (2011). Differential effects of perturbation direction and magnitude on the neural processing of voice pitch feedback. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 122*(5), 951–7. doi:10.1016/j.clinph.2010.08.010

Liu, Y., Hu, H., Jones, J. A., Guo, Z., Li, W., Chen, X., … Liu, H. (2015). Selective and divided attention modulates auditory-vocal integration in the processing of pitch feedback errors. *European Journal of Neuroscience, 42*(3), 1895–1904. doi:10.1111/ejn.12949

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. doi:DOI 10.1016/j.jneumeth.2007.03.024

Molholm, S., Martinez, A., Ritter, W., Javitt, D. C., & Foxe, J. J. (2005). The neural circuitry of pre-attentive auditory change-detection: An fMRI study of pitch and duration mismatch negativity generators. *Cerebral Cortex, 15*(5), 545–551. doi:10.1093/cercor/bhh155

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica, 42*(4), 313–329. doi:10.1016/0001-6918(78)90006-9

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544–2590. doi:10.1016/j.clinph.2007.04.026

Nakamura, K., Dehaene, S., Jobert, A., Le Bihan, D., & Kouider, S. (2007). Task-specific change of unconscious neural priming in the cerebral language network. *Proceedings of the National Academy of Sciences of the United States of America, 104*(49), 19643–19648. doi:10.1073/pnas.0704487104

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). Fieldtrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience, 2011*(2011). doi:doi:10.1155/2011/156869

Scheerer, N. E., Behich, J., Liu, H., & Jones, J. A. (2013). ERP correlates of the magnitude of pitch errors detected in the human voice. *Neuroscience, 240*, 176–85. doi:10.1016/j.neuroscience.2013.02.054

Scheerer, N. E., Liu, H., & Jones, J. A. (2013). The developmental trajectory of vocal and event-related potential responses to frequency-altered auditory feedback. *The European Journal of Neuroscience, 38*(8), 3189–200. doi:10.1111/ejn.12301

Segaert, K., Weber, K., de Lange, F. P., Petersson, K. M., & Hagoort, P. (2013). The suppression of repetition enhancement: A review of fMRI studies. *Neuropsychologia, 51*(1), 59–66. doi:10.1016/j.neuropsychologia.2012.11.006

Stolk, A., Todorovic, A., Schoffelen, J. M., & Oostenveld, R. (2013). Online and offline tools for head movement compensation in MEG. *NeuroImage, 68*, 39–48. doi:10.1016/j.neuroimage.2012.11.047

Tourville, J. A., Cai, S., & Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech (pp. 060180–060180). doi:10.1121/1.4800684

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981. doi:10.1080/01690960903498424

Toyomura, A., Koyama, S., Miyamaoto, T., Terao, A., Omori, T., Murohashi, H., & Kuriki, S. (2007). Neural correlates of auditory feedback control in human. *Neuroscience, 146*(2), 499–503. doi:10.1016/j.neuroscience.2007.02.023

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nat Neurosci, 3*(Suppl), 1212–1217.

Wolpert, D., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), 1880–1882.

Zarate, J. M., Wood, S., & Zatorre, R. J. (2010). Neural networks involved in voluntary and involuntary vocal pitch regulation in experienced singers. *Neuropsychologia, 48*(2), 607–18. doi:10.1016/j.neuropsychologia.2009.10.025

Zarate, J. M., & Zatorre, R. J. (2005). Neural substrates governing audiovocal integration for vocal pitch regulation in singing. *Annals of the New York Academy of Sciences, 1060*, 404–8. doi:10.1196/annals.1360.058

Zatorre, R., Evans, A., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256*(5058), 846–849. doi:10.1126/science.1589767

Zheng, Z. Z., Munhall, K. G., & Johnsrude, I. S. (2010). Functional Overlap between Regions Involved in

Speech Perception and in Monitoring One's Own Voice during Speech Production. *Journal of Cognitive Neuroscience, 22*(8), 1770–1781. Retrieved from <Go to ISI>://000279057700011

Zheng, Z. Z., Vicente-Grabovetsky, A., MacDonald, E. N., Munhall, K. G., Cusack, R., & Johnsrude, I. S. (2013). Multivoxel patterns reveal functionally differentiated networks underlying auditory feedback processing of speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33*, 4339–48. doi:10.1523/JNEUROSCI.6319-11.2013

6

AUDIOVISUAL RECALIBRATION OF VOWEL CATEGORIES

ABSTRACT

One of the most daunting tasks of a listener is to map a continuous auditory stream onto known speech sound categories and lexical items. A major issue with this mapping problem is the variability in the acoustic realizations of sound categories, both within and across speakers. Past research has suggested listeners may use visual information (e.g., lip-reading) to calibrate these speech categories to the current speaker. Previous studies have focused on audiovisual recalibration of consonant categories. The present study explores whether vowel categorization, which is known to show less sharply defined category boundaries, also benefit from visual cues. Participants were exposed to videos of a speaker pronouncing one out of two vowels, paired with audio that was ambiguous between the two vowels. After exposure, it was found that participants had recalibrated their vowel categories. In addition, individual variability in audiovisual recalibration is discussed. It is suggested that listeners' category sharpness may be related to the weight they assign to visual information in audiovisual speech perception. Specifically, listeners with less sharp categories assign more weight to visual information during audiovisual speech recognition.

This chapter is based on:

Franken, M. K., Eisner, F., Schoffelen, J.-M., Acheson, D. J., Hagoort, P., & McQueen, J. M. (2017). Audiovisual recalibration of vowel categories. In Proceedings of InterSpeech, Stockholm, Sweden.

6.1 INTRODUCTION

Speech perception is a remarkably complex skill. One of the most obvious issues every listener has to deal with is the enormous amount of variability in the speech signal. Acoustic variability in the speech signal is due to various factors, including the phonological context, the speaker's mood, speaker idiosyncrasies, the speaker's accent or dialect, etc.

One way in which listeners can deal with this variability in the acoustic signal is by recalibrating speech sound categories using additional sources of information (Nearey, 1989). For example, the so-called Ganong effect shows that listeners may use lexical information to bias speech perception (Ganong & F., 1980). A number of studies exposed listeners to a series of words where one consonant was replaced with an ambiguous sound. For example, Dutch listeners were presented with words like <witlo?>, where <?> represents a sound ambiguous between /f/ and /s/, making the word ambiguous between <witlof> ("chicory") and <witlos> (a pseudoword). The results showed that after exposure to a series of these items (where an /f/ interpretation would yield a known lexical item while the alternative /s/ interpretation would yield a pseudoword), listeners were biased to interpret the ambiguous sound as /f/, also subsequently in pseudowords, showing that they used lexical information to recalibrate speech categories (Eisner & McQueen, 2005; Norris, McQueen, & Cutler, 2003).

Another example is so-called audiovisual recalibration. This refers to listeners using visual (e.g. lip-reading) information to recalibrate speech perception. In these studies, listeners are exposed to videos of a speaker pronouncing a series of pseudowords. While the audio included an ambiguous consonant (for example, /a?a/, ambiguous between /aba/ and /ada/), the video was unambiguous in showing the speaker articulating either /aba/ or /ada/. Thus the visual information biased towards one of the two possible interpretations (for example, visual lip closure would bias towards /aba/). The results suggested that listeners had recalibrated their speech categories in a subsequent audio-only labeling task. Therefore, audiovisual information also may lead to recalibration of speech categories (Bertelson, Vroomen,

& De Gelder, 2003; Ley et al., 2012). A recent study compared lexical and audiovisual recalibration (van Linden & Vroomen, 2007).

The present study adds to the literature by investigating audiovisual recalibration in vowel categories. While all studies mentioned above have focused on audiovisual recalibration of consonant categories, this study investigates whether similar results hold for vowel categories. With respect to lexically-guided recalibration, recent studies have already shown that it occurs with vowel categories (Chládková, Podlipský, & Chionidou, 2017; Maye, Aslin, & Tanenhaus, 2008; McQueen & Mitterer, 2005). In the present paper, we investigate whether this holds for audiovisual recalibration as well. In addition, it is well established that vowel categories are less sharply defined and may be less stable compared to consonant categories (Kuhl, 1991). One may wonder whether the stability of phoneme categories (i.e., the "sharpness" of the phoneme boundaries) may affect this recalibration. If so, this may lead to two contrastive hypotheses. If the phoneme boundary is less sharp, this means the category is less clearly defined, and hence it could be more open to moving around. This would suggest listeners with less sharp boundaries show stronger recalibration effects. On the other hand, one could argue that if one has a sharp boundary and there is visual information that the boundary needs to move, one may be more likely to move it. This then would lead to stronger recalibration for listeners with sharper boundaries.

6.2 METHODS

6.2.1 Participants

10 native Dutch participants (7 females; age: M = 23 (SD = 4.06)) were recruited and provided informed consent according to the declaration of Helsinki. Participants were randomly assigned to one of two participant groups (5 participants in each group).

6.2.2 Materials

A 22-step vowel continuum was created using Praat (Boersma & Weenink, 2013). An original recording of the Dutch vowel /e:/ (spoken in context as /kapek/) was chosen and its source signal was extracted using linear predictive coding (LPC) and inverse filtering. The filter was manipulated by decreasing both F2 and F3 in 22 steps. The filter was then recombined with the source signal. The resulting vowels were recombined with the phonological context /kap_k/, resulting in a /kapek/-/kapøk/ continuum. So the whole continuum was created by manipulating F2 and F3 from a single / kapek/ recording. Both endpoints of the continuum are nonwords in Dutch.

Video stimuli were created by pairing each step of the audio vowel continuum with a video of the speaker's mouth articulating either /kapek/ (where the critical second vowel, /e/, is unrounded) or /kapøk/ (where the critical second vowel, /ø/, is rounded and hence visually distinct from /e/). This was the same speaker as in the auditory stimuli. Catch trial videos were created by adding a white dot (appearing for one frame only) in the middle of the video.

6.2.3 Paradigm

After instructions, participants were presented with a calibration block (audio only stimuli), a pre-test block (audio only stimuli), and then three to six (5 participants in each case) cycles alternating between exposure blocks (audiovisual stimuli) and post-test blocks (audio only stimuli).

In the calibration block, 12 steps along the continuum were presented (each 10 times) with a randomized order in a 2-alternative forced choice (2AFC) classification task. Participants were required to classify the stimulus as either /kapek/ or /kapøk/ by button press. For every participant individually, the most ambiguous step on the continuum (step x) and two neighboring steps (step x-2, x+2) were used in the remainder of the experiment. Step x-2 is closer to the /e/, and step x+2 is closer to /ø/.

In the pre-test, the same 2AFC classification task was used, this time including only the three most ambiguous steps (x, x-2, x+2). Each was presented 20 times, in randomized order.

In each exposure block, 20 videos were presented in a between-participant design. Videos were selected for each participant based on the participant group and their calibration phase data. For the /e/ group, videos of /e/ articulation (/kapek/) were paired with the audio of the most ambiguous step (step x), while videos of an /ø/ articulation (/kapøk/) were paired with an unambiguous /ø/ audio (step 22). In the /ø/ group, the /ø/ video was paired with the ambiguous audio (step x), while the /e/ video was paired with unambiguous /e/ audio (step 1). In every 20-trial block, two videos were catch videos (with a white dot). Participants were instructed to press a button as soon as they detected the white dot, in order to make sure they were looking at the videos.

The post-tests were similar in stimuli and task to the pre-test, but consisted only of 6 trials each (or 12 for half of the participants, but only the first 6 were analyzed). Exposure and post-test blocks alternated.

6.2.4 Analysis

All analyses were performed in R (R Core Team, 2013). For the calibration phase, a logistic regression was applied to determine the most ambiguous step along the continuum for every participant. This was defined as the step closest to the 50% cut-off of the logistic regression curve (i.e., the point along the continuum that would be classified by the participant as /ø/ in 50% of the cases). This step and two nearby steps (steps x, x-2, x+2) were used in the other blocks of the experiment.

For the pre-test and the post-tests, a generalized (binomial) linear mixed model was fitted to the data using a Laplace approximation with the R 'lme4' package (Bates, Mächler, Bolker, & Walker, 2015). Post-hoc investigation of the interaction term was performed using Holm's method for multiple comparison correction.

Finally, in order to take a closer look at individual variability, we compared the steepness of each participant's logistic curve fitted to their calibration phase data (i.e., the sharpness of the phonemic category boundary) with that participant's learning effect. The latter was quantified as the absolute value of the by-participant random slope coefficients for Time (pre- vs. post-test) from the generalized mixed model fitted to the data from pre- and post-tests.

6.3 RESULTS

Figure 6.1 shows the results over participants in the calibration block. The most ambiguous steps ranged across participants from step 6 to step 10. Based on the logistic regression, a logistic curve was fitted for each participant, as shown in Figure 6.1. These logistic curves vary across participants by two parameters, describing the boundary location (the point where the curve crosses the 0.50 line) and the boundary sharpness (the steepness of the curve).

Subsequently, the average percentages of /ø/ responses were calculated as a function of stimulus, test time (pre-test vs. post-test) and participant group. Figure 6.2 shows the results. Overall, the graph shows that most /ø/ percentages for the /e/ group (top row, /e/ group) decrease from pre-test to post-test, whereas this is not the case for the /ø/ group (bottom row, /ø/

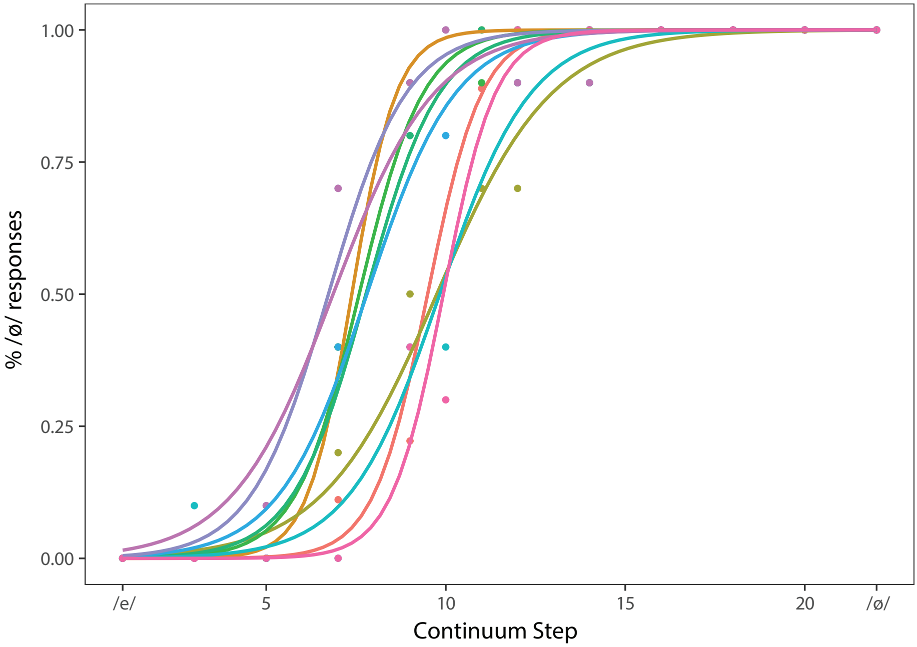

Fig. 6.1. Calibration results across participants. Colors represent participants; lines represent logistic regression fitted curves.

group). The clearest effects can be seen for stimulus step x (i.e., the most ambiguous vowel for each participant).

In order to explore the effects of the exposure to the video clips, we fitted a generalized mixed effects model to the data with fixed effects Time (pre vs. post), Group, and Stimulus, as well as their interaction terms, and a random intercept for participants and by-participants random slopes for Time. Analysis of the fixed effects estimates of the model shows significant main effects of Time (z = -2.87, p = .004), Group (z = 2.05, p = .040) and Stimulus (z = 8.48, p < .001), as well as a significant interaction between Group and Time (z = 2.26, p = .024). This suggests that the training affected the two participant groups differently. A closer analysis of the interaction effect revealed that while there was no significant difference between the pre-test and post-test for the /ø/ group (χ^2(1) = .23, p = .63), the /e/ group did categorize stimuli significantly less often as /ø/ in the post-test compared to the pre-test (χ^2(1) = 8.24, p = .0082), as was predicted.

The secondary aim of this study was to look at whether boundary sharpness was associated with the amount of audiovisual recalibration. Two alternative hypotheses were put forward. On the one hand, less

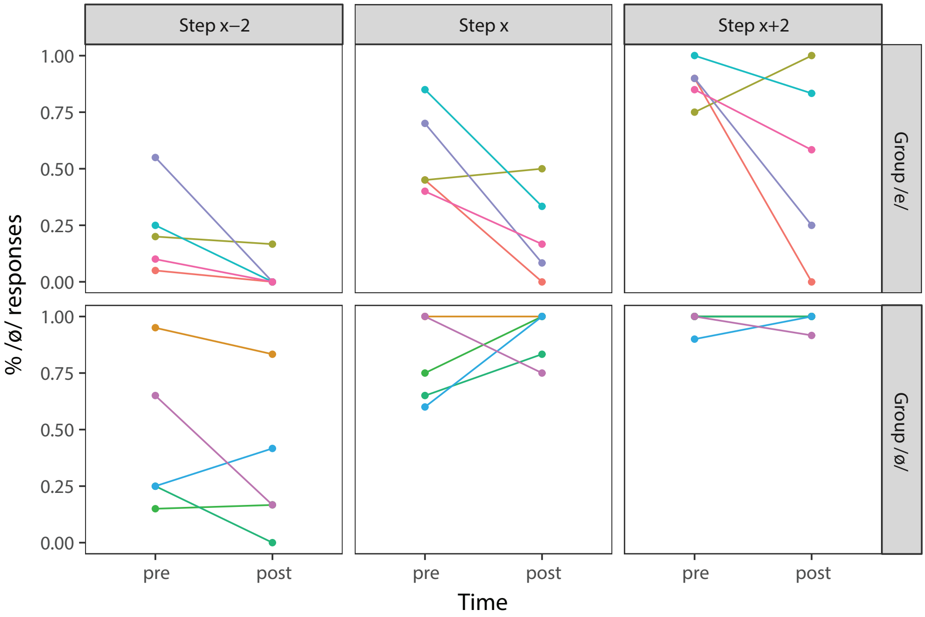

Fig. 6.2. Percentage /ø/ responses as a function of participant, participant group, stimulus step, and test time.

sharp boundaries could be freer to move around and therefore show more susceptibility to recalibration. On the other hand, less sharp boundaries might stay fuzzy under adaptation conditions and sharper boundaries might be more prone to recalibration. The boundary sharpness for each participant (steepness of the curves in Figure 6.1) tended to be associated with the participant's learning effect. Specifically, as the boundary steepness decreased, the learning effect increased (Figure 6.3), which is in line with the first hypothesis. However, caution is warranted for this interpretation, given this association was not significant (Pearson's $r(8) = -.60$, $p = .069$; Spearman's $\rho(8) = -.53$, $p = .12$).

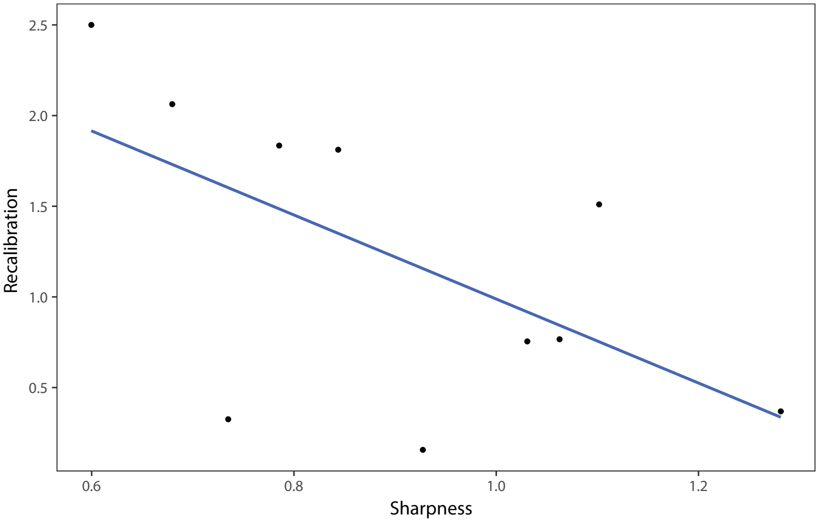

Fig. 6.3. Individual amount of audiovisual recalibration as a function of phoneme boundary sharpness.

6.4 DISCUSSION

In this study, we investigated whether listeners would recalibrate vowel categories using audiovisual information. The results suggest that this is indeed the case: The audiovisual perceptual learning block led to different effects in the two participant groups. Specifically, the group that was exposed to /e/ videos paired with ambiguous audio showed a reduction in /ø/

responses after perceptual learning.

These results suggest that participants recalibrated their vowel categories by shifting the /e/-/ø/ phoneme boundary. Recalibration of phoneme categories is one way by which listeners can attempt to solve the mapping problem between the incoming acoustic signal and abstract phonological categories. This is in line with what previous studies showed with audiovisual recalibration in consonant categories (Bertelson et al., 2003), with lexically-guided recalibration in vowel categories (Chládková et al., 2017; Maye et al., 2008; McQueen & Mitterer, 2005), and with the broader literature of what is known as cross-modal recalibration, for example in spatial cognition (Bedford, 1995).

The current data do not make it possible to investigate the temporal development of audiovisual recalibration. Previous studies on audiovisual recalibration in consonants have suggested that recalibration is a short-lived effect. A comparison between audiovisual and lexical recalibration suggested that audiovisual recalibration effects lived only for up to five test tokens after exposure (van Linden & Vroomen, 2007). The current study does not allow us to see whether this also holds for vowels, or whether the effect would last longer or shorter, given the less sharply defined phonemic boundaries in vowels than in consonants.

Closer investigation of the interaction between Time and Group suggested that only the /e/ group showed a difference between pre- and post-tests. The lack of audiovisual recalibration in the /ø/ group was unexpected. Inspection of individual participants' results showed that there was quite some variability across participants, also within the /ø/ group. Interestingly, the results in Figure 6.2 show that for step x in the /ø/ group, three out of five participants did show a change from pre- to post-test in the expected direction (an increase), while the two participants that did not show this effect already were at 100% in the pre-test (so they could not have shown any increase). This suggests that the lack of a group-level recalibration for the /ø/ group is a ceiling effect: there simply was no way to show recalibration. The 100% /ø/ responses in the pre-test suggest that the calibration block at least for these two participants did not adequately estimate the phoneme boundary. Moreover, also the other participants tended to show pre-test % /ø/ responses of over 50%, suggesting that step x may have been less ambiguous than was thought. One reason for this may be the asymmetry in the calibration stimulus materials. From Figure 6.1, it can be seen that all participants had their boundary in the left half of the continuum, and none of them categorized one of the four most /ø/-like stimuli (steps 16, 18, 20, 22) less than 100% as /ø/. This asymmetry may lead to a bias in participants' response patterns, as unbalanced exposure to two categories (e.g., more exposure to /ø/ stimuli compared to /e/) may bias subsequent categorization of these categories (as in selective adaptation (Eimas & Corbit, 1973; Kleinschmidt & Jaeger, 2016)) and therefore a mis-estimation of the boundary location. Future research should try to control for this bias, for example by excluding stimuli that don't differentiate between participants.

Finally, a closer analysis of individual variability in the data suggested that the amount of recalibration may be associated with phoneme boundary sharpness (although this result definitely needs replication). Specifically, the present findings indicate that well-defined categories are more robust to audiovisual recalibration, whereas less sharply defined categories are more susceptible to be recalibrated through audiovisual integration. Assuming this result shows up more robustly in a better-powered study, this suggests that listeners with fuzzy category boundaries assign more weight to visual information during audiovisual speech recognition. This may be explained as participants with less well-defined boundaries may find stimuli straddling the category boundary to be more ambiguous and therefore these participants stand to gain more from visual information compared to the participants with sharper category boundaries.

This study shows that listeners use visual information to recalibrate their vowel categories. This is in line with past research on consonants and lexically-guided recalibration, but extends it to vowel categories, which are known to have less sharply defined categorical boundaries. Moreover, although the current data do not warrant any strong conclusions about individual differences, it was suggested that individuals with fuzzy or less sharp perceptual category boundaries assign more weight to visual information during audiovisual speech recognition and therefore show increased audiovisual recalibration. If this finding is corroborated, given that vowel categories have less sharp boundaries compared to consonants, there ought to be audiovisual recalibration for vowel categories, given consonants have shown audiovisual recalibration in previous research. This is indeed what was found in the current study.

REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). doi:10.18637/jss.v067.i01

Bedford, F. L. (1995). Constraints on perceptual learning: objects and dimensions. *Cognition, 54*(3), 253–297. doi:10.1016/0010-0277(94)00637-Z

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science, 14*(6), 592–7. doi:10.1046/J.0956-7976.2003.PSCI_1470.X

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

Chládková, K., Podlipský, V. J., & Chionidou, A. (2017). Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance, 43*(2), 414–427. doi:10.1037/xhp0000333

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*(1), 99–109. doi:10.1016/0010-0285(73)90006-6

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224–238. doi:10.3758/BF03206487

Ganong, W. F., & F., W. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110–125. doi:10.1037/0096-1523.6.1.110

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review, 23*(3), 678–691. doi:10.3758/s13423-015-0943-z

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107. doi:10.3758/BF03212211

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., & Formisano, E. (2012). Learning of New Sound Categories Shapes Neural Response Patterns in Human Auditory Cortex. *Journal of Neuroscience, 32*(38), 13273–13280. doi:10.1523/JNEUROSCI.0584-12.2012

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543–562. doi:10.1080/03640210802035357

McQueen, J. M., & Mitterer, H. (2005). Lexically-driven perceptual adjustments of vowel categories. *ISCA Workshop on Plasticity in Speech Perception*, (June), 233–236.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America, 85*(5), 2088. doi:10.1121/1.397861

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

R Core Team. (2013). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statictical Computing. Retrieved from http://www.r-project.org

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1483–1494. doi:10.1037/0096-1523.33.6.1483

7

IMPLICIT PHONETIC RECALIBRATION AND ITS INFLUENCE ON SPEECH ADAPTATION

ABSTRACT

The relationship between speech perception and production is a debated issue in the speech sciences. A recent study (Lametti, Krol, Shiller, & Ostry, 2014) suggested that explicit perceptual learning affects subsequent speech motor adaptation. The current study was set up to examine whether this relationship also holds for implicit perceptual learning. Participants were exposed to audiovisual speech samples, where the audio was ambiguous between /e/ and /ø/, while the video disambiguated the vowel in a between-participants audiovisual recalibration paradigm. This led participants that were exposed to /e/ videos matched with the ambiguous vowel to shift their phoneme boundary towards /ø/. Subsequently, we tested whether this boundary shift would affect speech motor learning in an altered auditory feedback paradigm. If the boundary has been shifted by perceptual recalibration towards the vowel that is produced, the altered feedback would bring the perceived vowel closer to the phoneme boundary, which should lead to more speech adaptation. Alternatively, if the boundary has been shifted away from the vowel being produced, this should lead to less speech adaptation. In contrast to the findings reported by Lametti et al., these hypotheses were not borne out. It is suggested that implicit perceptual learning may be talker- and stimulus-specific, and therefore does not generalize to speech production learning.

7.1 INTRODUCTION

The relation between speech perception and speech production is a hotly debated issue in the speech sciences. What is clear is that the relations between perceptual representations, acoustics, and articulation are complex. However, changes in speech perception, which is remarkably plastic, seem to have little direct impact on speech production. The current study examines the link between speech perception and speech production by examining whether implicit perceptual learning transfers to subsequent speech motor learning.

The plasticity of speech perception is clear from studies on lexically-guided perceptual learning (Eisner & McQueen, 2005; Kraljic, Brennan, & Samuel, 2008; Norris, McQueen, & Cutler, 2003). For example, Norris et al. exposed Dutch listeners to /s/- or /f/-final Dutch words such as radijs, 'radish', or witlof, 'chicory', where the final fricative was an ambiguous sound between /f/ and /s/. Subsequently, listeners altered their perceptual categorization of an [f]-[s] continuum depending on the lexical context during exposure: Listeners who heard the ambiguous fricative in a context where /f/ was expected identified it as /f/, whereas listeners who had heard it in /s/-biased contexts treated it as /s/. This suggests that listeners adapted their categories by shifting the phoneme boundary in the appropriate direction. Similarly, when exposed to audiovisual stimuli, listeners may use cues from lip-reading to adapt perceptual categories (Bertelson, Vroomen, & De Gelder, 2003; Ley et al., 2012; van Linden & Vroomen, 2007). In Bertelson et al., for example, listeners were exposed to videos of a speaker articulating /aba/ or /ada/, paired with audio that was ambiguous between the two. Subsequent identification tasks of the ambiguous auditory stimulus indicated that listeners had used the visual cues to adapt their auditory categories by shifting the phoneme boundary in the appropriate direction.

The phenomena of lexically and audiovisually guided recalibration show that speech perception is very plastic. However, changes in perception seem to have little immediate impact on speech production. For example, listeners quickly come to understand speech in a foreign accent, yet they don’t change their own articulation as a consequence. A recent study by Lametti et al. (2014), however, suggested that perceptual changes may affect speech production in the context of speech motor learning. In this study, participants first performed an identification task on a head-had continuum, while they received explicit feedback. In other words, participants were explicitly told whether their identification response was ‘correct’. Different groups of participants were given different feedback on what was correct, which led to shifts in phonetic identification either towards “head” (i.e., fewer “head” responses) or towards “had” (i.e., more “head” responses). Subsequently, participants performed a speech adaptation task, where they articulated ‘head’, while auditory feedback was being manipulated in real time (the vowel in ‘head’ was shifted towards ‘hid’). The results showed that perceptual learning had a clear impact on subsequent speech motor learning. Specifically, if perceptual learning led to a boundary shift towards ‘head’, participants compensated more strongly for the shift due to auditory feedback. This is in line with earlier work involving altered auditory feedback, which showed that participants compensate more or more quickly when the feedback perturbation shifted a vowel towards a nearby phoneme boundary rather than towards the vowel distribution’s centroid (Niziolek, Nagarajan, & Houde, 2013).

The current study also investigated whether a perceptual change affects speech motor learning, and used a paradigm similar to the one used by Lametti et al. (2014). However, the perceptual task Lametti et al. used (explicit feedback in an identification task) may not be optimal. Although the perceptual learning task worked (as shown by the effect on subsequent motor adaptation), the presence of explicit feedback could invite decision-level strategies on the part of the participant. In everyday speech, listeners use perceptual learning to cope with the huge amount of variability in the speech signal. Importantly, in these cases perceptual learning usually occurs implicitly, without conscious attention of the listener and without explicit feedback. Therefore, one may wonder whether Lametti et al.’s results, which show transfer from an explicit perceptual learning task to speech adaptation, generalize to an implicit perceptual learning task. In addition, little is known about how participants' phoneme representations are altered by the explicit perceptual task Lametti et al. used.

Another task that has been claimed to lead to a shift in phoneme boundaries is the perceptual recalibration task discussed earlier. The advantage of such tasks over that used by Lametti et al. is the absence of explicit feedback, as in natural perceptual learning. Several studies have suggested the recalibration effect is truly perceptual, rather than involving a decision bias or occurring at a post-perceptual stage (Clarke-Davidson, Luce, & Sawusch, 2008; Keetels, Pecoraro, & Vroomen, 2015; McQueen, Cutler, & Norris, 2006; Mitterer & Reinisch, 2013). The current study used an implicit audiovisual recalibration paradigm in the style of Bertelson et al. (2003) to investigate perceptual learning influences on subsequent speech adaptation. Audiovisual rather than lexical recalibration was used, as van Linden & Vroomen (2007) have suggested that the former leads to larger recalibration effects. Note that all studies that have made use of the audiovisual recalibration paradigm introduced by Bertelson et al. (2003) to date have focused on consonants. In a pilot study (Franken et al., 2017, see also chapter 6), however, we have suggested that vowels show audiovisual recalibration as well.

The main research question in the current study is this: If participants show implicit perceptual learning, does this affect subsequent speech adaptation? As in the Lametti et al. study, after perceptual learning, participants were exposed to altered auditory feedback in a speech adaptation paradigm, in order to examine whether perceptual learning would lead to a group difference in speech adaptation to altered auditory feedback. Specifically, participants were exposed to audiovisual speech samples that paired an ambiguous auditory vowel with an unambiguous video of either /e/ or /ø/, in a between-participant design. Perceptual learning was measured by auditory-only identification tasks both before and after audiovisual exposure. We expected participants exposed to /e/ videos to identify the ambiguous vowel more often as /e/ after audiovisual exposure, while the participants exposed to /ø/ videos were expected to identify the ambiguous vowel more often as /ø/. In addition, participants read aloud visually presented Dutch

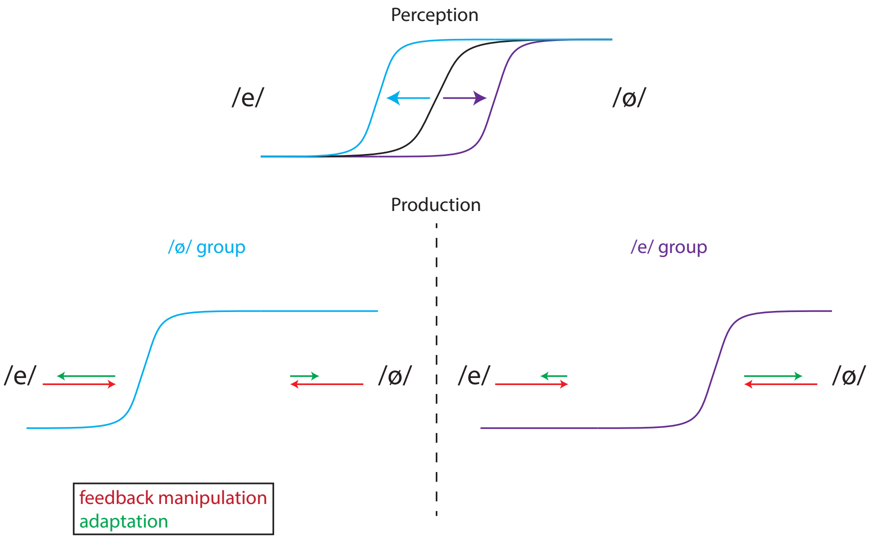

Fig. 7.1. Illustration of the main hypothesis. Top: the perception task (audiovisual recalibration) leads to a shift of the phoneme boundary on the /e/-/ø/ vowel continuum. The /ø/ group shifts the phoneme boundary towards the /e/ side of the continuum (light blue), while the /e/ group shifts the phoneme boundary towards the /ø/ side of the continuum (purple). Bottom: illustration of the hypothesized effects of the perceptual task on production for the /ø/ group (left, blue) and the /e/ group (right, purple). Red arrows indicate the shift introduced by the F2 perturbation of auditory feedback, green arrows indicate participants' response to the altered feedback. If the perturbation is towards a close-by phoneme boundary, there is a strong compensatory response (long green arrow), if the perturbation is towards a far-away phoneme boundary, there is a weak compensatory response (short green arrow).

words containing the vowels /e/ and /ø/, while they were exposed to altered auditory feedback, both before any perceptual test and after audiovisual exposure. Auditory feedback was perturbed by shifting the vowel towards the /e/-/ø/ phoneme boundary. Figure 7.1 illustrates the hypothesized effects of audiovisual recalibration on subsequent speech adaptation. We expected that the closer the phoneme boundary is to the vowel participants perceive through altered auditory feedback, the stronger participants would adapt to compensate for the feedback perturbation. Therefore, a phoneme boundary shift towards the /ø/ end of the vowel continuum, as is expected for the participants exposed to the /e/ videos (Figure 7.1, top), is hypothesized to induce stronger adaptation when articulating /ø/, but less strong adaptation when articulating /e/ (Figure 7.1, bottom right). As the boundary should be shifted in the opposite direction in the other group of participants (who were exposed to /ø/ videos), we expected effects in the opposite direction on the production side: weaker adaptation when articulating /ø/, but stronger when articulating /e/ (Figure 7.1, bottom left).

7. 2 METHODS

7.2.1 Participants

Thirty healthy volunteers (age: M = 21, SD = 2.1; 26 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local ethics board committee (CMO Arnhem / Nijmegen). All participants had normal hearing, were native speakers of Dutch and had no history of speech and/or language pathology. Participants were randomly assigned to one of two participant groups (fifteen participants in each group).

7.2.2 Stimuli

A 22-step vowel continuum was created using Praat (Boersma & Weenink, 2013). An original recording (stereo channels, sampling rate = 48kHz) of the Dutch vowel /e:/, spoken by a single male native speaker of Dutch, was chosen and its source signal was extracted using linear predictive coding (LPC) and inverse filtering. The filter was manipulated by decreasing both F2 and F3 in 22 steps, and recombined with the source signal. The resulting vowels were embedded in the phonological context /kap_k/, resulting in a /kapek/-/kapøk/ continuum. A pilot study (Franken et al., 2017, see also chapter 6) using the same stimuli suggested there may be a bias towards /e/ responses during identification of the entire continuum (compared to just the three most ambiguous stimuli). In an effort to eliminate this effect, it was decided to restrict the continuum to the first 15 steps (starting from the /e/ end of the continuum), resulting in a 15-step /kapek/-/kapøk/ continuum.

For the audiovisual stimuli, videos of the same speaker were recorded with a JVC GZ-MG135 digital video camera at 25 frames per second. Video stimuli were created by pairing videos of the same speaker's mouth articulating either /kapek/ or /kapøk/ (still images in Figure 7.2) with each step of the acoustic /kapek/-/kapøk/ continuum. For each vowel step, there were two different videos in order to reduce potential effects of peculiarities of any one video. In the videos, only the mouth of the speaker was visible, from the chin up to the nose (Figure 7.2). Catch trial videos were created by adding a white dot (appearing for the duration of one video frame only) in the middle of the video.

In the production tasks, stimuli were presentations of one of the Dutch words STEEN (/sten/, 'stone'), BEEK (/bek/, 'stream'), LEEK (/lek/, 'layman'), STEUN (/støn/, 'support'), BEUK (/bøk/, 'beech') or LEUK (/løk/, 'fun'). The words were presented in capital letters at the center of a computer screen for 1.5s.

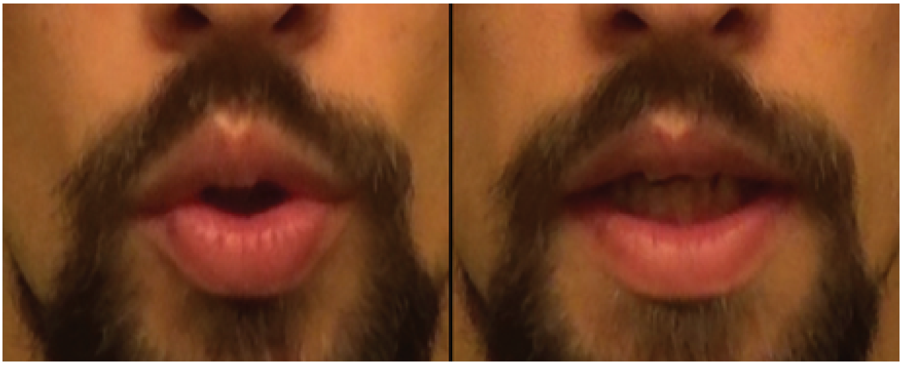

Fig. 7.2. Still images from two videos. On the left, a still from an /ø/ video, and on the right a still from an /e/ video. Both stills are taken at the center of the vowel in question.

7.2.3 Procedure

The experimental paradigm consisted of 7 blocks, illustrated in Figure 7.3. First, there was a production baseline block ("Baseline"), followed by two perceptual blocks ("Calibration" and "Perceptual pre-test"). Then, the main part of the experiment consisted of cycles of blocks with audiovisual exposure, perceptual post-tests, and production post-tests. The purpose of the Calibration block prior to the main experiment was to determine for every participant individually the most ambiguous point on a /e/-/ø/ continuum. To this end, participants heard one out of nine continuum steps (steps 1, 3, 5, 7, 8, 9, 11, 13 and 15 of the 15-step continuum) and were instructed to identify the stimulus in a two-alternative forced choice task.

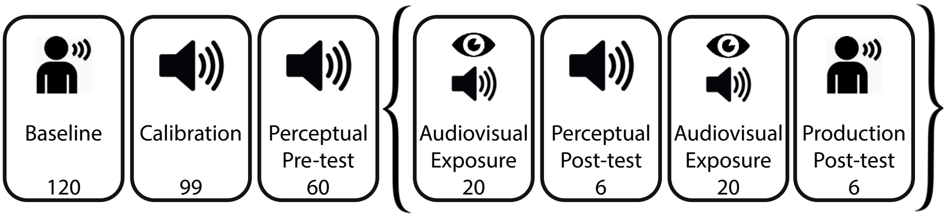

Fig. 7.3. Overview of experimental paradigm. Numbers at the bottom of each experimental block indicates the number of trials in the block. The block sequence between the curly brackets was repeated nine times.

This was done for 99 trials, of which the first 9 were excluded for analysis (leading to 10 repetitions of each stimulus being used for calibration). After the calibration block, a logistic regression was fitted to the participant's identification responses, to determine the step in the vowel continuum that was closest to 50% cutoff of the psychometric curve. This step (step X) and two neighboring steps at two step sizes distance (steps x-2, x+2) were selected for the remainder of the experiment. Subsequently, in a perceptual pre-test, participants listened to tokens of these three stimuli (steps x-2, x, x+2) and performed the same two-alternative forced choice identification task. This pre-test consisted of 60 trials (20 repetitions of each stimulus).

In the audiovisual exposure block, participants were exposed to the video stimuli in a between-participant design. At this point, participants were randomly assigned to one of two participant groups. Participants in the /e/ group were exposed to videos of a /kapek/ articulation paired with ambiguous audio (that participant's step x). The videos were contrasted with videos of a /kapøk/ articulation paired with unambiguous /kapøk/ audio (step 22). Participants in the /ø/ group were exposed to /kapøk/ videos paired with ambiguous (step x) audio, contrasted with /kapek/ videos with unambiguous /kapek/ audio (step 1). There were two versions of each video, so four different videos for each participant. These were presented each five times in randomized order (20 trials in total). The audiovisual exposure block was followed by a short perceptual post-test, which was the same as the perceptual pre-test, with now only two repetitions of each stimulus (6 trials).

The other blocks of the experiment measured speech production. The experiment started with a baseline production block, which consisted of a classic speech adaptation task (Houde & Jordan, 1998; Purcell & Munhall, 2006). On each of the 120 trials, a single word was displayed on the screen, to be read aloud by the participant. The participant's speech was recorded and played back to them in real time. Over the first 30 trials, feedback was not manipulated. In trials 31 to 90, F2 in every vowel was manipulated. The F2 in /e/ vowels was decreased by 10%, the F2 in /ø/ vowels was increased by 10%. In the last 30 trials, auditory feedback was not manipulated. Finally, after the audiovisual exposure blocks, six production trials were administered. Participants were instructed again to read aloud the visually presented words, while auditory feedback was manipulated in the same way as in trials 31-90 of the baseline block.

The last four experimental blocks (audiovisual exposure, perceptual post-test, audiovisual exposure, production post-test) was repeated as a sequence 9 times, amounting to a total of 39 blocks.

The experiments were written in Matlab, using the Psychophysics extension (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). Auditory feedback was manipulated with Audapter (Cai, Boucek, Ghosh, Guenther, & Perkell, 2008; Tourville, Cai, & Guenther, 2013) using a MOTU MicroBook II audio interface. Auditory stimuli were delivered through Sennheiser HD-202 headphones, and speech was recorded with a Sennheiser ME62 omnidirectional condenser microphone in a sound proof booth.

7.2.4 Analyses

All analyses were performed with R (R Core Team, 2013). For the calibration block, a logistic regression was applied to determine, for each participant, the most ambiguous step in the vowel continuum. This was defined as the step closest to the 50% cut-off of the fitted logistic regression curve (i.e., the point along the continuum that would be classified by the participant as /ø/ in 50% of the cases). This step and two nearby steps (separated by one step, so eventually steps x, x-2, x+2) were used in the remaining auditory and audiovisual blocks of the experiment.

For the perceptual pre- and post-tests, a generalized (binomial) linear mixed effects model was fitted to the data using a Laplace approximation with the R 'lme4' package (Bates, Mächler, Bolker, & Walker, 2015). All p values reported for fixed effects are based on Satterthwaite's approximations of the degrees of freedom (Kuznetsova, Brockhoff, & Christensen, 2016). Post-hoc investigations of interaction terms were done using Holm's method for multiple comparison correction. Finally, we had a closer look at the individual variability of the perceptual data by correlating the steepness of each participant's logistic curve fitted to the calibration data (i.e., the sharpness of the phoneme category boundary) with that participant's learning effect. The latter was quantified as follows. For both groups separately, a linear mixed effects model was run. Subsequently the participant-specific coefficients for the post-test effect (i.e., the effect of the post-test relative to the pre-test) were taken as proxies for the amount of perceptual learning. The sign of the coefficients was reversed for the participants of the /e/ group, such that, for participants in both groups, larger values indicated stronger perceptual learning in the hypothesized direction.

For the production data, for every produced word, the average F2 value was extracted and converted to the bark scale with the following equation:

$$F2_{bark}=13 \cdot tan^{-1}(0.00076 \cdot F2_{Hz})+3.5 \cdot tan^{-1}((F2_{Hz}/ 7500)^{2})$$

where $F2_{\mathrm{bark}}$ refers to the F2 value in bark, and $F2_{\mathrm{Hz}}$ to the F2 value in Hertz. Subsequently, outliers were identified by calculating, for each participant and each vowel separately, a z-score for both F1 and F2 values for every trial, and identifying the trials where the F1 or F2 z-score exceeded -3 or +3. These values were excluded from further analyses. For statistical inference, F2 values were entered in linear mixed effects models. Regressors for Vowel (/e/ vs. /ø/) and Block (baseline vs. post-test) were contrast-coded (values =0.5 vs. -0.5). The regressor Feedback, corresponding to the three phases within the baseline block, was coded as two dummy variables representing the contrasts perturbation – start (coded as [-0.5, 0.5, 0] for the factor levels start, perturbed, end) and end – start (coded as [-0.5, 0, 0.5]). All p values reported for fixed effects reflect Satterthwaite's approximations of the degrees of freedom.

7. 3 RESULTS

7.3.1 Perception

For the calibration block, a logistic regression was performed for the results of every participant separately. The resulting fitted curves are shown in Figure 7.4. As is clear from the figure, participants varied both in location of their phoneme boundary (taken as the point in the continuum where the fitted curve crosses the 50% cut-off), as well as in the slope of the curve.

The main analysis of interest with respect to the perceptual part of the experiment focuses on the effect of the audiovisual exposure on perceptual identification. The results are presented in Figure 7.5. From this figure, two main conclusions can be drawn. First, looking at the results from the post-test (i.e., after audiovisual exposure, solid lines in Figure 7.5), there is a clear difference between the two groups. This indicates that audiovisual exposure indeed affected subsequent speech perception. Specifically, percentage /ø/ responses were lower for the /e/ group compared to the /ø/ group. This was expected, as in the /e/ group, the ambiguous audio was associated with /e/ videos (leading to a decrease in /ø/ responses), whereas in the /ø/ group, the ambiguous audio was paired with /ø/ videos (leading to an increase in /ø/ videos). Second, if, however, we also take into account the results from the perceptual pre-test (the identification task before audiovisual recalibration, dotted lines in Figure 7.5), it seems the difference between the two groups was mainly driven by the /e/ group. A generalized (binomial) linear mixed model was fitted to the responses with fixed main effects for Block (pre-test vs. post-test), Stimulus, and Group, as well as all two- and three-way interactions. The random effect structure included random intercepts for participants and by-participant random slopes for Block, Stimulus, and a

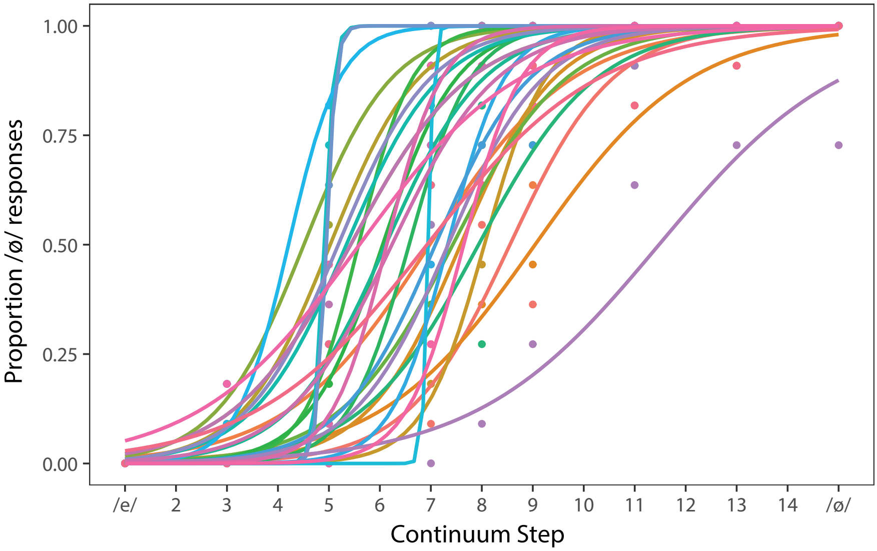

Fig. 7.4. Results for the calibration block. Results are presented as proportion of /ø/ identifications as a function of continuum step. Continuum steps are indicated along the x axis, ranging from /e/ to /ø/. Solid lines represent the fitted curves from the logistic regression, one for each participant.

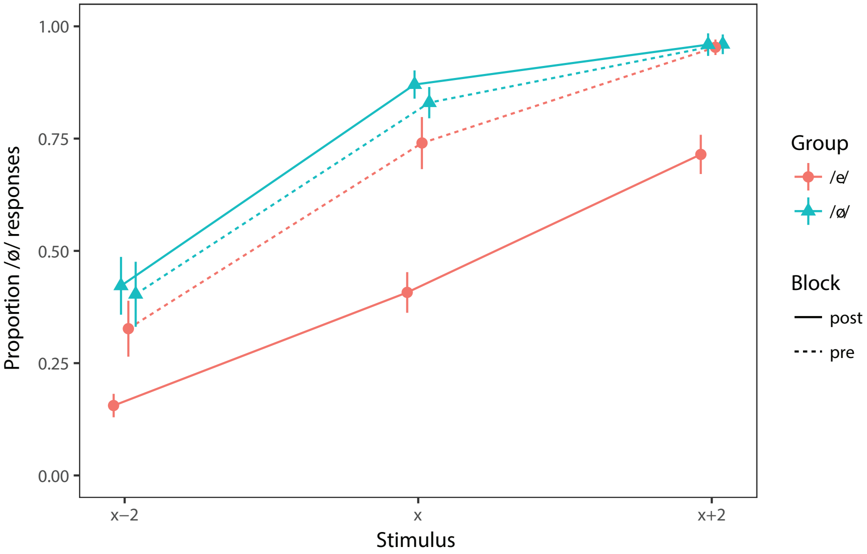

Fig. 7.5. Group-level results of perceptual pre- and post-tests.

Table 7.1. Fixed effects results for perception task.

	Estimate	Std. Error	z value	*p*
(Intercept)	1.44	.35	4.05	<.001*
/ø/ Group	.52	.50	1.05	.30
Block post	-1.86	.38	-4.95	<.001*
Stimulus	1.24	.16	7.72	<.001*
/ø/ Group : Block post	2.01	.55	3.69	<.001*
/ø/ Group : Stimulus	.04	.22	.19	.85
Block post : Stimulus	-.49	.14	-3.51	<.001*
/ø/ Group : Block post : Stimulus	.44	.21	2.10	.036*

Block:Stimulus interaction. The results showed significant main effects for Block and Stimulus, and significant interactions between Group and Block, Block and Stimulus, as well as a significant three-way Block:Group:Stimulus interaction (Table 7.1). Closer investigation of the Group-by-Block interaction revealed a significant change from pre- to post-test in the /e/ group ($\chi^2(1) = 24.54$, $p < 0.001$, Holm-corrected), but not in the /ø/ group ($\chi^2(1) = 0.14$, *n.s.*).

The absence of a perceptual learning effect for the /ø/ group was unexpected. It may be due to a ceiling effect: Figure 7.5 shows quite high percentages of /ø/ responses in the pre-test, even for step x, which was the most ambiguous point along the vowel continuum in the calibration block. With such high percentages already in the pre-test, there is little room for evidence of perceptual learning, given that we would expect an increase for the /ø/ group (as indeed observed, at least numerically). This argument is supported two additional analyses. First, we find that the recalibration effect decreases over time for the /e/ group (β = .21 (std. err. = .045), $z = 4.70$, $p < .001$), but not for the /ø/ group (β = -.038 (.055), $z = -.70$, $p = .49$), in line with previous findings of dissipating effect sizes (van Linden & Vroomen, 2007). Every post-test consisted of 6 trials. Figure 7.6 shows the average proportion of /ø/ responses as a function of trial number, showing that the effect for the /e/ group is strongest at the first trials. Second, for the /ø/ group, we find a negative correlation between the random by-participant intercepts (the reference level for Block was the pre-test, so this reflects variability in the pre-test responses) and the random by-participant slopes

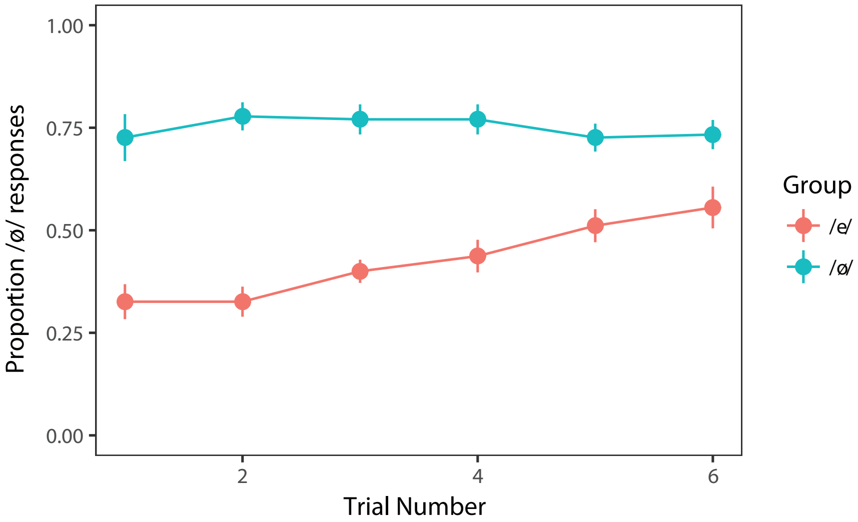

Fig. 7.6. Proportion of /ø/ responses as a function of Trial Number (averaged across post-tests). The effect gradually diminishes over time in the /e/ group, but not in the /ø/ group. Error bars indicate the standard error.

for Block (Spearman's ρ = -0.61, S = 902, p = 0.02). In other words, the higher the probability that participants identified stimuli in the pre-test as /ø/, the less likely they showed an increase in the post-test, and thus evidence of perceptual learning.

Another source of individual variability is the nature of the phoneme boundary. In a pilot study of the perceptual part of the current study (Franken et al., 2017), we suggested listeners with less sharp boundaries may show more perceptual learning. The slopes of the fitted curves for the calibration phase (Figure 7.4) were correlated with the amount of perceptual learning. Although for the current data, there was no significant correlation (one-tailed Spearman's *rho* = 0.053, S = 4258, *n.s.*), we found a trend in the expected direction when we restrict the analysis to the /e/ group (one-tailed Spearman's *rho* = -0.27, S = 710, p = 0.17). This suggests that, at least for the /e/ group, a less sharp boundary may be associated with stronger perceptual learning.

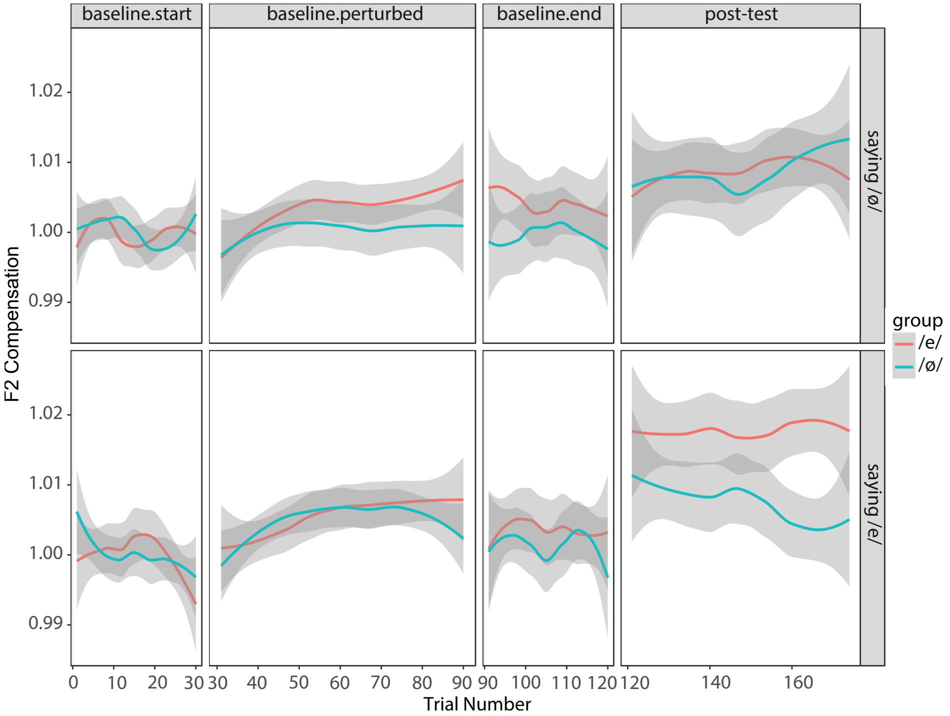

Fig. 7.7. Overview of F2 compensation is shown as a function of trial number, participant group, block, and vowel produced. F2 values were divided by the average in baseline.start, and normalized for perturbation direction (higher numbers indicate compensation). The colors indicate participant group, columns indicate experimental phase (the baseline block is divided in baseline.start, where feedback was normal, baseline.perturbed, where feedback was perturbed, and baseline.end, where feedback was normal again), rows indicate the vowel produced. Lines are generated by a loess smoothing function.

7.3.2 Production

It was investigated whether and how the participants responded to the F2 perturbations in the auditory feedback during speech production. All production analyses focused on changes in participants’ F2 as a function of the feedback (see Figure 7.7).

First, a look at the baseline block should indicate whether participants responded to the altered auditory feedback, independent of any perceptual learning. A linear mixed effects model was run on the F2 values (in bark), with fixed main effects for Vowel (the vowel contained in the stimulus) and Feedback phase (start, perturbed or end phase within the baseline block), as well as their interaction. The random effects structure included by-participant random intercepts and slopes for Vowel, Feedback, and their

Table 7.2. Fixed effects results for production task.

	Estimate	Std. Error	df	t value	p
(Intercept)	13.12	.12	22.37	114.00	< .001*
Vowel	1.42	0.15	5.45	9.70	< .001*
Feedback_pert-start	.026	.014	29.01	1.92	.065$^{(*)}$
Feedback_end-start	-.0050	.019	28.95	-.26	.80
Vowel:Feedback_pert-start	.086	.026	28.88	3.38	.0021*
Vowel:Feedback_end-start	.0086	.030	28.82	.29	.77

interaction, as well as by-item random intercepts. The results are shown in Table 7.2. The F2 values differed between the two vowels (t(5.45) = 9.70, p < .001), and the difference between the perturbed and the initial unperturbed (start) trials was almost significant (t(29.01) = 1.92, p = .065). A significant interaction was found between Vowel and the Feedback phase (t(28.88) = 3.38, p = .0021), suggesting participants adjusted their F2 in response to the perturbation differently for the two vowels. Pairwise post-hoc tests were carried out to investigate this interaction further. The perturbation trials showed higher F2 values than the start trials for the /e/ vowel (value = -.069, χ^2(1) = 8.16, p = .026, Holm-corrected), but not for the /ø/ vowel (value = .022, χ^2(1) = 1.45, p = .23, uncorrected). For the /e/ vowel, the F2 values in the perturbation trials were also almost higher than in the end trials (value = .035, χ^2(1) = 5.44, p = .098, Holm-corrected). None of the other comparisons were significant, suggesting that participants did compensate for alterations in the feedback for the /e/ vowel, but not for the /ø/ vowel. In addition, participants showed no statistical evidence of aftereffects in the end phase.

A second set of analyses examined how responses to the altered auditory feedback were affected by the intervening audiovisual exposure. To that end, F2 values in the perturbed trials of the baseline block were compared to F2 values in the production post-tests (Figure 7.8). For each vowel separately, a linear mixed effects model was run with main effects for Block (baseline vs. post-test), Group, and their interaction. Random effects were included as by-participant random intercepts and random slopes for Block, as well as by-item random intercepts. For the vowel /e/, the results showed an overall increase in F2 from baseline to post-test (t(28.01) = 2.97, p = .006), and an

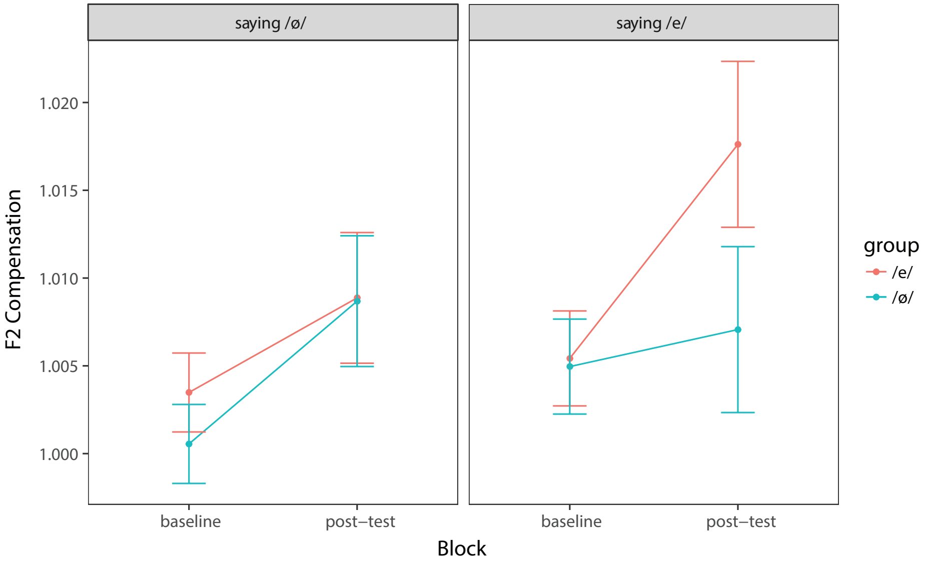

Fig. 7.8. F2 compensation shown as a function of experimental block, participant group, and vowel produced. F2 values were divided by the average in baseline.start, and normalized for perturbation direction (higher numbers indicate compensation). It can be seen that there is an overall, group-independent increase in compensation from baseline to post-test when saying /ø/, but only the /e/ group showed an increase in compensation in the post-test when saying /e/; this is the only group difference in the data.

interaction between Block and Group (*t*(28.01) = -2.09, *p* = .046). A closer analysis of this interaction showed that the /e/ Group increased F2 in the post-test relative to the baseline block (value = .17, $\chi^2(1) = 12.80$, *p* < .001, Holm-corrected), while there was no change for the /ø/ group (Figure 7.8, right column). For the /ø/ vowel, there was no significant interaction. There was however an overall decrease in F2 from baseline to post-test (*t*(27.92) = -2.91, *p* = .007), suggesting, unlike in the analysis for the /e/ vowel, a group-independent increase in compensation (Figure 7.8, left column).

7. 4 DISCUSSION

In the current study, we investigated whether implicit perceptual learning through audiovisual recalibration affected subsequent speech adaptation. With respect to the perceptual task, the results show that, after exposure to videos where an ambiguous vowel was paired with an unambiguous /e/ video, listeners were less likely to categorize ambiguous auditory stimuli as /ø/. This suggests that listeners used visual information (lip-reading) to recalibrate vowel categories by shifting the phoneme boundary. This is in line with earlier studies on audiovisual recalibration of consonant categories (Bertelson et al., 2003), lexical recalibration with consonants (Norris et al., 2003) and vowels (Chládková, Podlipský, & Chionidou, 2017; McQueen & Mitterer, 2005), as well as with our pilot study using audiovisual recalibration with vowels (Franken et al., 2017, see also chapter 6).

Although the group difference in the perceptual post-test revealed a striking effect of audiovisual exposure, the comparison between pre-test and post-test results (Figure 7.3) did not show a group-level recalibration in the /ø/ group. This was unexpected, though note that previous research often did not collect pre-test data and therefore did not report such pre- vs. post-test comparisons. The current results suggest the lack of recalibration in the /ø/ group may be due to a ceiling effect. As the probability of an /ø/ response was already over 80% in the pre-test, there was little room for audiovisual recalibration to increase the percentage of /ø/ responses (although there was a small trend in that direction). In fact, the magnitude of audiovisual recalibration was associated with listeners' performance in the pre-test: the higher the percentage of /ø/ responses in the pre-test, the less audiovisual recalibration was observed in the /ø/ group. This is in line with a ceiling effect, and suggests that participants already changed their responses between the calibration and pre-test blocks. Note that both groups showed high pre-test scores, so this is not something peculiar about the participants in the /ø/ group. A possible reason for the difference between calibration and pre-test is the difference in stimulus range, as that was the only difference between the two blocks (stimuli ranged across the whole continuum in calibration vs. only three tokens centered at the most ambiguous token in the pre-test). In the absence of an unambiguous /e/, maybe everything tends to sound more like /ø/. Previous studies have shown that the stimulus range can indeed affect categorization results (Repp & Liberman, 1987). Categorization of vowels is possibly more sensitive than consonants to the range of presented stimuli, given that vowel perception is known to be less categorical compared to consonants (Fry, Abramson, Eimas, & Liberman, 1962). Therefore, we suggest it is important to include a pre-test when studying audiovisual recalibration.

With regards to the main research question, whether perceptual recalibration affected subsequent speech adaptation, the current results do not support a direct transfer from implicit perceptual learning to speech motor learning. Specifically, when the amount of compensation for feedback alterations before and after audiovisual exposure were compared (Figure 7.8), the only significant group difference was an increase in compensation when producing /e/ for the /e/ group but not for the /ø/ group. The two groups showed an equivalent change in F2 when producing /ø/. The group difference for /e/ was opposite to what was expected based on the Lametti et al. (2014) study. Given the /e/ group shifted their phoneme boundary towards the /ø/ end of the vowel continuum during perceptual learning (i.e., fewer /ø/ responses; Figure 7.5), the phoneme category boundary was farther away from their /e/ productions, so speakers should be less likely to compensate. In contrast, the /ø/ group as a whole did not show audiovisual recalibration, though some individuals showed a trend towards recalibration by shifting the boundary towards the /e/ end of the continuum. So when producing /e/ they should, if anything, have been more likely to compensate compared to the /e/ group. However, the opposite was found for the /e/ productions, while there was no group difference whatsoever for the /ø/ productions.

It is unclear how this unexpected pattern of results can be explained. We tentatively propose one explanation in light of the distributional learning account of perceptual learning proposed by Kleinschmidt & Jaeger (2016). Their account proposes that perceptual learning (including recalibration and selective adaptation) is sensitive to both the mean and the variance of a phonetic distribution. Therefore, we tentatively propose the hypothesis that selective adaptation, or continued exposure to prototypical examples of a particular phoneme (Eimas & Corbit, 1973; Vroomen, van Linden, de Gelder, & Bertelson, 2007), would reduce the variance of that phoneme's distribution and thus lead to increased adaptation under altered feedback, because of a narrower target representation.

To our knowledge, this idea of selective adaptation leading to increased speech motor adaptation has not been tested and thus requires further investigation. If this hypothesis is correct, however, it could explain part of the current results. During audiovisual exposure, the /e/ group was exposed to prototypical examples of /ø/, and the /ø/ group to prototypical examples of /e/, but also possibly to prototypical examples of /ø/. This is because, given the ceiling effect in the /ø/ group due to the high pre-test scores of percentage /ø/ responses, the 'ambiguous' vowel may have been closer to prototypical /ø/, especially when paired with /ø/ videos. This exposure to reduced-variance /ø/ in both groups could potentially explain the increased compensation when saying /ø/ that was indeed observed in both groups. However, in principle it would also predict increased compensation for the /ø/ group when saying /e/, which we did not find. This may potentially be explained in line with findings in visual selective adaptation (Chopin & Mamassian, 2012), suggesting that the effect of selective adaptation is stronger when the distribution deviates more from the expected distribution. As the vowel /ø/ is much less frequent than /e/ in Dutch, selective adaptation may be more pronounced for /ø/. Finally, increased compensation was observed for the /e/ group when saying /e/. Although this group was exposed to ambiguous (or even rather /ø/-like) vowels, after audiovisual exposure a strong recalibration effect was observed. This may have led the participants in the /e/ group to perceive the audiovisual samples as relatively clear instances of /e/, inducing selective adaptation which would yield an effect opposite to that based on recalibration. Note that the recalibration effect in the perceptual post-tests in fact did decrease over time, which may suggest an added effect of selective adaptation.

An alternative, simpler explanation involves disregarding the production results for the /ø/ group, as they did not show audiovisual recalibration, and disregarding the results for the production of /ø/ vowel, as it didn't show compensation even in the baseline production block. What is left is an increase in compensation for the production of /e/ by the /e/ Group. As this effect is opposite to what was expected based on audiovisual recalibration, it may be due to selective adaptation. If either this or the more complex explanation is correct, this would suggest selective adaptation does, but recalibration does not, transfer to adaptation in speech production. Additional studies are necessary to examine the validity of these explanations. A previous study on audiovisual recalibration with consonant categories has in fact shown that over time, recalibration can turn into selective adaptation (Vroomen et al., 2007). These authors showed that with repeated exposure to ambiguous stimuli, participants initially showed a recalibration effect and over time start showing selective adaptation instead. In other words, the direction of the perceptual change flips to the opposite direction. This confirms that both processes could be active at the same time.

The current results suggest that perceptual learning in the current study did not generalize to speech adaptation in the way that was hypothesized. This is in contrast with the study by Lametti et al. (2014), who did find clear generalization of perceptual learning to speech adaptation. There were some differences between their paradigm and the experimental design in the current study. With respect to the production task, while speakers in the Lametti study produced head while the F1 in the auditory feedback was manipulated, speakers in the current study produced words containing either /e/ or /ø/, while F2 was being manipulated in different directions (upwards for /ø/ and downwards for /e/). In other words, in the current study, speakers had to adapt to two feedback manipulations (in opposite directions) that were alternated, which may have made it harder to adapt adequately. Note, however, that a previous study showed that speakers can quite readily modify their speech movements to correct for multiple auditory transformations simultaneously (Rochet-Capellan & Ostry, 2011). Note, in addition, that we did observe evidence of speech-motor adaptation, so alternating between opposite feedback shifts did not block all adaptation.

The main difference between the current study and the Lametti et al. (2014) study is the perceptual task that was used. While participants in the Lametti et al. study categorized a vowel continuum while they received explicit feedback, participants in the current experiment were exposed to an implicit audiovisual recalibration paradigm. While both studies showed perceptual learning effects in a subsequent auditory-only categorization task, the explicitness of the perceptual training paradigm may be an important factor. While there is little previous literature on the categorization with feedback task that speaks to the locus of perceptual learning, the literature on perceptual recalibration (be it lexically or visually guided) has shown that this type of implicit perceptual learning may generalize to other segments (Kraljic & Samuel, 2006), other syllable positions (Jesse & McQueen, 2011) or other phonetic contexts (Mitterer, Chen, & Zhou, 2011). However, a study by Reinisch et al. (2014) showed no evidence of generalization to the same phoneme contrast cued differently or to another phoneme contrast with the same cues, suggesting that recalibration may take place at context-dependent sub-lexical units. A recent study (Mitterer, Scharenborg, & McQueen, 2013) suggested that generalization of recalibration may depend on the nature of the input. The authors showed that recalibration did not generalize to different allophones of the same phonemes, suggesting that recalibration was stimulus-specific. In addition, Eisner & McQueen have shown that lexical recalibration did not generalize to a different talker (Eisner & McQueen, 2005). This is in line with other studies (Clarke-Davidson et al., 2008; Keetels et al., 2015) suggesting that both lexical and audiovisual recalibration take place at an early perceptual level, rather than inducing a decision bias. In contrast, providing explicit feedback may have forced a change in explicit vowel categories at a later cognitive stage, possibly involving a decision bias.

In the context of these prior results, the current study suggests that there are constraints on the generalization of perceptual learning to speech motor learning. While Lametti et al. (2014) have shown perceptual effects on speech adaptation, these results may be dependent on an explicit perceptual learning paradigm. The implicit recalibration task used in the current study did not lead to similar generalization effects, although it did lead to some change in the production (be it in the opposite direction). This suggests that in the current study, perceptual learning is affecting production, but only in an indirect way. One should consider that participants in the current study did a perceptual learning task with stimuli recorded by one talker, using a low-variability stimulus set, while the production task of course involved responding to real-time manipulations of their own voice. If perceptual recalibration is talker- or stimulus-specific, the perceptual learning effect should indeed not affect subsequent production directly. On the other hand, we have tentatively argued that the current paradigm may have given rise, over time, to selective adaptation effects, which did affect speech production. The recalibration may be stimulus-specific, and therefore does not affect production immediately, but through our paradigm selective adaptation may arise, which does percolate through to production.

REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). doi:10.18637/jss.v067.i01

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science, 14*(6), 592–7. doi:10.1046/J.0956-7976.2003. PSCI_1470.X

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433–436.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In *Proceedings of the 8th Intl. Seminar on Speech Production* (pp. 65–68). Strasbourg, France.

Chládková, K., Podlipský, V. J., & Chionidou, A. (2017). Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance, 43*(2), 414–427. doi:10.1037/xhp0000333

Chopin, A., & Mamassian, P. (2012). Predictive properties of visual adaptation. *Current Biology, 22*(7), 622–626. doi:10.1016/j.cub.2012.02.021

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. doi:10.3758/PP.70.4.604

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*(1), 99–109. doi:10.1016/0010-0285(73)90006-6

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224–238. doi:10.3758/BF03206487

Franken, M. K., Eisner, F., Schoffelen, J.-M., Acheson, D. J., Hagoort, P., & McQueen, J. M. (2017). Audiovisual recalibration of vowel categories. In *Proceedings of InterSpeech*. Stockholm, Sweden.

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. Retrieved from <Go to ISI>://WOS:A-1962CAZ3900001

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213–1216. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9469813

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review, 18*(5), 943–950. doi:10.3758/s13423-011-0129-2

Keetels, M., Pecoraro, M., & Vroomen, J. (2015). Recalibration of auditory phonemes by lipread speech is ear-specific. *Cognition, 141*, 121–126. doi:10.1016/j.cognition.2015.04.019

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3. *Perception, 3*, 36(ECVP Abstract Supplement).

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review, 23*(3), 678–691. doi:10.3758/s13423-015-0943-z

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*(1), 54–81. doi:10.1016/j.cognition.2007.07.013

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R Package version 2.0-33. Retrieved from https://cran.r-project.org/package=lmerTest

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014). Brief periods of auditory perceptual training can determine the sensory targets of speech motor learning. *Psychological Science, 25*(7), 1325–36. doi:10.1177/0956797614529978

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., & Formisano, E. (2012). Learning of New Sound Categories Shapes Neural Response Patterns in Human Auditory Cortex. *Journal of Neuroscience, 32*(38), 13273–13280. doi:10.1523/JNEUROSCI.0584-12.2012

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*, 1113–1126. doi:10.1207/s15516709cog0000

McQueen, J. M., & Mitterer, H. (2005). Lexically-driven perceptual adjustments of vowel categories. *ISCA Workshop on Plasticity in Speech Perception*, (June), 233–236.

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science, 35*(1), 184–197. doi:10.1111/j.1551-6709.2010.01140.x

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language, 69*(4), 527–545. doi:10.1016/j.jml.2013.07.002

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition, 129*(2), 356–361. doi:10.1016/j.cognition.2013.07.011

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What Does Motor Efference Copy Represent? Evidence from Speech Production. *Journal of Neuroscience, 33*(41), 16110–16116. doi:Doi 10.1523/Jneurosci.2137-13.2013

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America, 120*(2), 966. doi:10.1121/1.2217714

R Core Team. (2013). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statictical Computing. Retrieved from http://www.r-project.org

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics, 45*(1), 91–105. doi:10.1016/j.wocn.2014.04.002

Repp, B. H., & Liberman, A. M. (1987). Phonetic Category Boundaries are Flexible. In S. Harnad (Ed.), *Categorical Perception* (pp. 89–112). New York, NY: Cambridge University Press.

Rochet-Capellan, A., & Ostry, D. J. (2011). Simultaneous Acquisition of Multiple Auditory-Motor Transformations in Speech. *Journal of Neuroscience, 31*(7), 2657–2662. doi:Doi 10.1523/Jneurosci.6020-10.2011

Tourville, J. A., Cai, S., & Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech (pp. 060180–060180). doi:10.1121/1.4800684

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1483–1494. doi:10.1037/0096-1523.33.6.1483

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory?visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45*(3), 572–577. doi:10.1016/j.neuropsychologia.2006.01.031

8

MODULATIONS OF THE AUDITORY M100 IN AN IMITATION TASK

ABSTRACT

Models of speech production explain event-related suppression of the auditory cortical response as reflecting a comparison between auditory predictions and feedback. The present MEG study was designed to test two predictions from this framework: 1) whether the reduced auditory response varies as a function of the mismatch between prediction and feedback; 2) whether individual variation in this response is predictive of speech-motor adaptation. Participants alternated between online imitation and listening tasks. In the imitation task, participants began each trial producing the same vowel (/e/) and subsequently listened to and imitated auditorily-presented vowels varying in acoustic distance from /e/. Results replicated suppression, with a smaller M100 during speaking than listening. Although we did not find unequivocal support for the first prediction, participants with less M100 suppression were better at the imitation task. These results are consistent with the enhancement of M100 serving as an error signal to drive subsequent speech-motor adaptation.

This chapter is based on:

Franken, M. K., Hagoort, P., & Acheson, D. J. (2015). Modulations of the auditory M100 in an imitation task, Brain and Language, 142, 18-23. doi: 10.1016/j.bandl.2015.01.001

8.1 INTRODUCTION

Feedback plays a crucial role in speech motor control as it signals a speaker whether a speech motor action was successful or not. In order to account for an extensive number of findings related to adaptation and feedback processing in motor control, a number of theories have posited a monitoring mechanism that utilizes forward models. Here, motor commands sent to speech articulators also send an ‘efference copy’ through forward models that predict the somatosensory and/or auditory consequences of those commands (Hickok, 2012; Houde & Nagarajan, 2011; Tian & Poeppel, 2010).

The workings of this internal forward model for speech production are thought to be reflected in a reduction of the auditory cortex response to self-produced speech relative to listening to recordings of the same speech. Magneto- and Electrophysiological studies have found a reduction of the M100, a well-known auditory component that occurs roughly 100ms after audio onset (Naatanen & Picton, 1987). Theoretical models (Guenther, Ghosh, & Tourville, 2006; Hickok, 2012; Houde & Nagarajan, 2011) explain M100 suppression (M100S) as reflecting a comparison mechanism: if the internal forward model’s prediction of the auditory consequences of speech commands matches the actual auditory input, the cortical response is attenuated. When the prediction does not match the auditory feedback entirely, there is a reduction of M100S (i.e., the auditory cortex shows less suppression). This reduction of M100S then acts as an error signal, driving compensatory mechanisms that adapt motor output towards internal, auditory goals.

Over the last decade, a number of properties of M100S have emerged. Using magnetoencephalography (MEG), Houde and colleagues showed that masking the auditory feedback abolished M100S (Houde, Nagarajan, Sekihara, & Merzenich, 2002). Subsequent studies have shown that the amount of suppression can be modulated by properties of the feedback (Behroozmand & Larson, 2011; Heinks-Maldonado, Nagarajan, & Houde, 2006; Ventura, Nagarajan, & Houde, 2009). Such feedback can also reflect self-produced variation (Sitek et al., 2013). For instance, Niziolek et al. (2013) found that the M100S correlates with the distance between a people's production and the centroid of their vowel space. Together, these studies support a view in which M100S reflects a match between predicted and actual auditory feedback. Additional support for this view comes from direct cortical recordings (Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Flinker et al., 2010), where it was also argued that the reduced M100S may be caused either by less neural suppression (i.e., less SIS) or via neural enhancement on top of stable SIS. This neural enhancement was hypothesized to reflect prediction error in the auditory processing, whereas SIS is helpful in distinguishing self-produced speech from external speech. However, few studies have directly linked M100S to behavioral output (Chang et al., 2013). The current study was designed to more clearly establish this link by having people engage in an imitation task in which auditory feedback is critical to performance, and in which motor consequences of mismatch are likely to adapt in real time.

The goal of this study was to replicate M100S in an imitation task and to test two claims of the aforementioned theories of speech motor control. First, if the M100S indexes the match between a prediction and the incoming auditory signal, then the amount of suppression should relate to the degree of mismatch between the prediction and the auditory signal (Behroozmand & Larson, 2011; Heinks-Maldonado et al., 2006; Houde et al., 2002; Liu, Meshman, Behroozmand, & Larson, 2011). Second, if a reduction of M100S serves as an error signal to drive motor adaptation, then individual variation in the amount of suppression should be predictive of the individual variation in imitation aptitude. Specifically, people who show a smaller M100S should show larger adaptations to their speech as they more readily produce error signals that could drive imitation performance.

To test these claims, we measured MEG during online imitation and listening tasks. In the speech imitation task, subjects were instructed to produce the vowel /e/ when a visual cue appeared. At the same moment, subjects heard a recording of themselves producing an auditory stimulus that was the same, close, or far from /e/, and they were asked to imitate this stimulus by adjusting their ongoing vowel production. By varying the acoustic distance between /e/ and the auditory target, we were able to investigate whether this acoustic distance modulated the magnitude of the M100S and people's subsequent speech-motor performance.

8.2 METHODS

8.2.1 Participants

Thirty-two healthy volunteers (age: M = 21.8, SD = 5.2; 21 females) participated after providing written informed consent in accordance with the Declaration of Helsinki and the local ethics board committee (CMO region Arnhem / Nijmegen). All subjects had normal hearing, normal or corrected-to-normal vision, were right-handed and were native speakers of Dutch.

8.2.2 Stimuli

Auditory stimuli included subject-specific vowel recordings made in a preceding session. Participants were recorded while producing the Dutch vowels /i/, /ɪ/, /e/, /ɛ/ and /a/. The vowels were equalized in length (to 1s), in pitch (to the subject's average pitch) and in intensity (to 82dB). All audio editing was done in Praat (Boersma & Weenink, 2013). Catch stimuli were made by shifting the pitch in the final 100ms of the sound up by 8mel. The mel scale is a psychoacoustic scale first proposed by Stevens, Volkmann & Newman (1937).

8.2.3 Experimental Design and Procedure

Subjects alternated between an imitation task and a listening task in a blocked design. Imitation trials started with a fixation cross (1260-2260ms, jittered). Participants were instructed to start saying /be/ as soon as three exclamation marks appeared. Participant's speech onset triggered playback (average delay = 51ms, sd = 4.3ms) of an auditory stimulus (9 x /i/, /ɪ/, /ɛ/, /a/ and 12 x /e/). The first block was a baseline block consisting of 72 /e/ trials. The participant's task was to imitate the auditory stimulus as accurately as possible by adapting his/her pronunciation of the vowel to match the auditory stimulus. The exclamation marks remained visible for 1500ms after speech onset. The inter-trial interval was 1000ms.

The listening trials were similar, except the stimulus played directly when the exclamation marks appeared, and participants did not have to vocalize. To maintain focus on the stimuli, participants performed an auditory detection task in which they pressed a button when they detected a rising pitch at the end of the stimulus. Four such trials were added to each block, and feedback for accuracy was provided after each catch trial. The experimental design was implemented using a PC running Presentation® software (Version 16.2, www.neurobs.com). Participants' speech was recorded with a single microphone in the magnetically shielded room with a sampling rate at 44100Hz and all auditory stimuli were presented binaurally via MEG-compatible air tubes. The complete experiment had 872 trials, which were randomized within blocks.

8.2.4 Data Acquisition

We used an MEG system (VSM/CTF systems, Port Coquitlam, Canada) with 275 axial gradiometers. Three localization coils, fixed to anatomical landmarks (nasion, left and right preauricular points), were used to determine head position relative to the gradiometers. Head position was monitored online by the experimenter and if necessary corrected between the experimental blocks (Stolk, Todorovic, Schoffelen, & Oostenveld, 2013).

All data were low-pass filtered by an anti-aliasing filter (300Hz cut-off), digitized at 1200Hz and stored for offline analysis. Participants were seated upright, with the head rested against the back of the helmet and touching the top of the helmet. A small cushion was used to fix the head's position so as to minimize free head movement.

8.2.5 Analyses

8.2.5.1 Behavioral

In the participants' speech recordings, speech onset served as a marker for the beginning of the epoch of interest. The first 900ms of each epoch were divided into 10 time bins of 90ms, for which the average F1 and F2 values were calculated using an iterative Burg algorithm. Formant values were expressed in the psychoacoustic bark scale (Zwicker, 1961). The acoustic distance to the stimulus (i.e. the imitation target) was quantified as the Euclidian distance in F1/F2 space between the subject's speech and the stimulus. Two-way repeated-measures ANOVAs with factors Time Bin and Vowel were carried out on the F1 data, the F2 data and the acoustic distance separately. One participant failed to perform the task appropriately and was excluded from all further analyses.

8.2.5.2 MEG

All MEG data analyses were performed using the Fieldtrip toolbox for EEG/MEG-analysis (Oostenveld, Fries, Maris, & Schoffelen, 2011). Data was resampled at 300Hz, and epochs of interest were defined from 1s before to 2s after stimulus onset. Trials were excluded if the range and variance of the MEG signal differed by at least an order of magnitude from the other trials of the same subject. Imitation trials in which subjects failed to speak or spoke too softly (therefore not triggering the stimulus playback) were discarded as well. On average, 7 trials per subject were rejected, resulting in an average of 827 trials per subject. The datasets for two subjects lacked one imitation block and one participant's dataset lacked both an imitation and a listening block. As these still contained 792 and 744 trials before trial rejection, we did not exclude their data.

For artifact removal, we performed an independent components analysis (ICA; see Makeig, Jung, Bell, Ghahremani, & Sejnowski, (1997)). Prior to ICA, a principal components analysis (PCA) was run on the data to reduce the number of dimensions to 100. Visual inspection of the ICA components' time courses and scalp topographies, resulted in rejection of an average of 13 components (SD = 6) per subject. Components were identified as containing artifacts when they showed clear signatures of eye movement, heartbeat (Vigario, Sarela, Jousmaki, Hamalainen, & Oja, 2000), or speech-related muscle artifacts. The latter type was identified as having peripheral topography (indicative of sources close to the jaw muscles) and a time course that correlated well with speech on- and offset (a step-like response from the moment of speech onset until speech offset). Subsequently, the data was transformed back to sensor space and band-pass filtered using a two-pass Butterworth filter (order: 4; pass-band: 1-12Hz).

Event-related fields (ERF) were calculated by averaging the trials time-locked to stimulus onset (baseline: 400-200ms before stimulus onset). All further analyses were done on the planar gradients (Bastiaansen & Knosche, 2000). Statistical testing was done within a time window of interest from 50ms to 250ms after stimulus onset. An ANOVA (factors Task and Vowel) was performed on the ERF data averaged across the time window. To select which channels to use, a data-driven approach was performed by selecting the channels that were among the 80 most active channels within the time window in both the listening and the speaking data for each subject (mean = 53; SD = 7).

8.2.5.3 Individual differences

The amount of M100S was calculated per condition as the difference between the ERF values for the listening and speaking trials, normalized by the ERF values in the listening data (i.e., (listening - speaking) / listening). The results were averaged across conditions. Imitation performance was quantified as the normalized change in acoustic distance to the stimulus: the distance between the participant's speech and the auditory stimulus was calculated for the first and the last time bin of the subject's speech. Performance was then calculated as the difference between the first time bin and the last bin per condition, normalized by the distance to target in the first time bin (i.e., (distance_to_targetbin1 – distance_to_targetbin10) / distance_to_targetbin1). Finally, the results were averaged across conditions. So, a value of 0 for behavioral performance indicates the subject are just as far from the target at the end as in the beginning of the trial, whereas a value of 1 indicates perfect imitation. An additional analysis was run to see whether the average amount of suppression was correlated with stimulus dispersion in vowel space. The latter was quantified as the average Euclidian distance to each participant's centroid, the point in vowel space defined by the average F1 and F2 values across all five stimuli.

8.3 RESULTS

8.3.1 Behavioral Performance

In order to assess whether participants appropriately performed the imitation task, an initial analysis focused on people's speech output as a function of time across each imitation condition (see Figure 8.1). Results show that for both the first formant (F1) and second formant (F2), participants started at similar values every trial, and the formant values subsequently diverged depending on the vowel. Results of a 9 (Time) X 5 (Vowel) repeated-measures ANOVA on F1 values (see Figure 8.1a) showed significant main effects of both Time ($F(9, 270) = 23.97$; $p < 0.0001$) and Vowel ($F(4, 120) = 145.8$; $p < 0.0001$), as well as a significant Time x Vowel interaction ($F(36, 1080) = 106.9$; $p < 0.0001$). Similarly, for F2 (Figure 8.1b) both main effects as well as the interaction were significant (Time: $F(9, 270) = 9.25$; $p < 0.0001$; Vowel: $F(4, 120) = 79.49$; $p < 0.0001$; interaction: $F(36, 1080) = 60.78$; $p < 0.0001$). With the exception of /ɪ/ condition, all vowels differed from the /e/ condition. Note that although all five vowels were phonemically distinct in Dutch, an important cue to distinguish /ɪ/ from /e/ is vowel duration, a parameter that was lost in this experiment (Adank, van Hout, & Smits, 2004). Importantly, however, these behavioral results demonstrate that participants did imitate the auditory stimuli.

In order to quantify whether the changes in F1 and F2 brought participants closer to the imitation target, we examined the Euclidian distance in F1-F2 space between subjects' speech output and their imitation target on a given trial (Figure 8.2c). Results showed significant main effects of Time ($F(9, 270) = 42.1$; $p < 0.0001$) and Vowel ($F(4, 120) = 45.79$; $p < 0.0001$) as well as an interaction ($F(36, 1080) = 34.62$; $p < 0.0001$). Post-hoc tests revealed

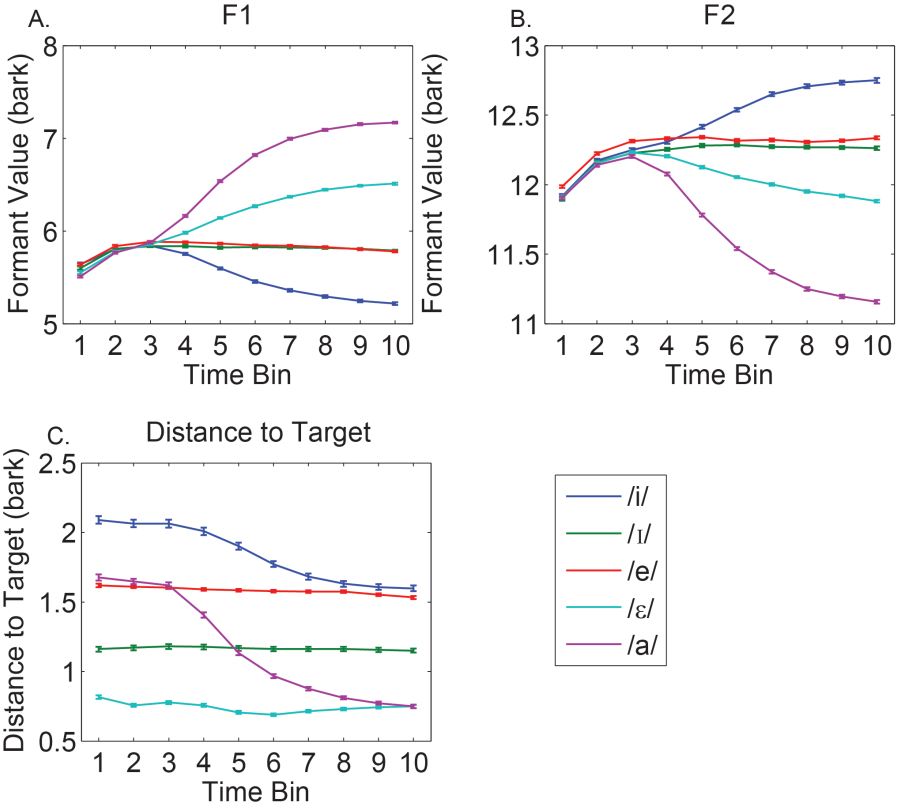

Fig. 8.1 Average behavioral results across participants. Formant values are expressed in bark (Zwicker, 1961). All error bars represent within-subject standard error of the mean A. F1 formant values averaged per time bin and vowel across subjects. Each time bin is 90ms long. B. F2 formant values averaged per time bin and condition across subjects. C. Distance to the imitation target averaged per time bin and condition across subjects.

that for the vowels /i/ and /a/ the distance to the target decreased between time bins 1 and 10 ($t(30) = 6.08$, $p < 0.0001$ and $t(30) = 10.99$, $p < 0.0001$). However, we did not find any effect for the other vowels (all *ps* > 0.5). This may be due to the large individual differences, non-linearities in vowel space, and/or over-adaptations, such as when participants initially get close to the target but then adapt too much, ending farther from the target at the end. Consistent with a previous imitation study (Kent, 1974), these over-adaptations were found for at least one vowel in 14 out of 31 subjects (/a/ - 1 subject; /ɛ/ - 10 subjects; /ɛ/ and /a/ - 3 subjects). No over-adaptations were found for /i/, /ɪ/ or /e/.

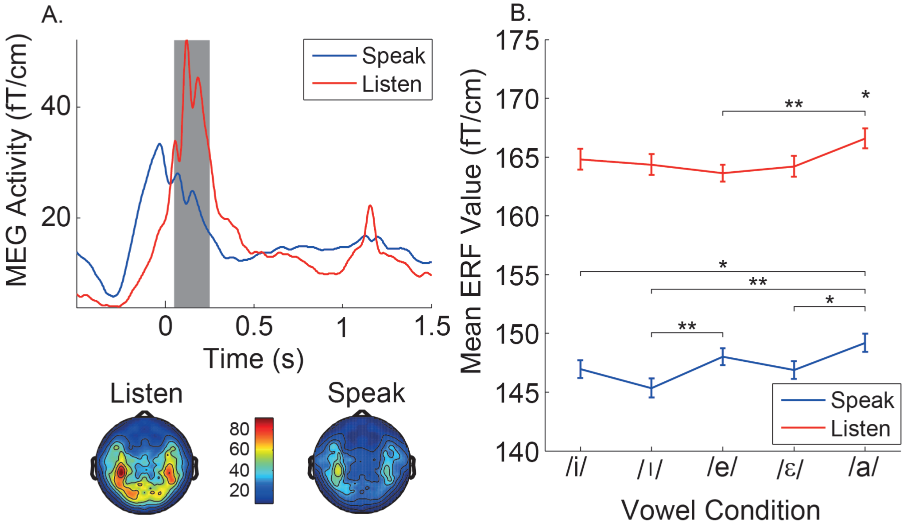

Fig. 8.2 ERF task effects at the group level. A. Group-level ERF values per task (all sensors), time-locked to auditory stimulus onset. The time window of interest (50-250ms) is indicated by the grey shading. Scalp topographies of the listening and speaking ERF (50- 250ms) are shown as well. B. Group-level ERF averages by Task and Vowel. ERF values were averaged over time (50-250ms) and channels (subject-specific channel selection, see methods). (*: significant only without multiple error correction; **: significant after Bonferroni correction). Error bars represent within-subject standard error of the mean.

8.3.2 Speaking-Induced suppression modulated at the group level

One of the main questions of the present study was whether we could replicate the M100S in an imitation task, and if so, whether the M100S was modulated by the acoustic distance of the stimuli from the starting vowel. Comparison of the ERF in the speaking and the listening task replicated M100S in the imitation paradigm (Figure 8.2a). A clear activity peak can be seen in the listening ERF in the time window (50-250ms) that corresponds to early auditory activity (i.e., the M100), and scalp topographies are consistent with auditory sources. In roughly the same time window, an activity peak is present in the speaking condition with similar topography. Crucially, the listening data clearly shows a larger peak compared to the speaking data. This result thus extends previous work which has observed SIS in simple speech production tasks and in altered feedback paradigms.

It is important to point out that activity was observed prior to stimulus onset (0 ms) in the speaking task (see Figure 8.2a). This pre-stimulus activity might reflect an auditory response to speech onset, or possibly speech motor activity, given that stimulus presentation in the speaking condition was triggered by a voice key (i.e., subjects started speaking already before stimulus onset; see methods). Visual inspection of the scalp topography of this early activity (not shown) was very similar to the post-stimulus activity. Consistent with the M100S post-stimulus, this pre-stimulus activity peak was still lower than the post-stimulus listening peak.

In order to examine how the different conditions modulated the ERFs, we calculated the average ERF per task and per vowel. A 2 (Task) X 5 (Vowel Condition) repeated-measures ANOVA on average ERFs revealed main effects of Task ($F(1, 30) = 85.58$; $p < 0.0001$) and Vowel Condition ($F(4, 120) = 5.72$; $p < 0.0001$), and a Task x Vowel Condition interaction that approached significance ($F(4, 120) = 2.05$; $p = 0.09$). In order to investigate the interaction more closely, the amount of suppression was calculated and compared between vowels (M100S = $(\mathrm{ERF}_{\mathrm{listen}} - \mathrm{ERF}_{\mathrm{speak}}) / \mathrm{ERF}_{\mathrm{listen}}$). We found only partial support for our prediction that the magnitude of M100S would scale with deviation from the initial speaking condition. In line with our predictions, the less deviant conditions (i.e. those acoustically closer to /e/), /ɪ/ (mean = 0.11, within-subject SE = 0.0039) and /ɛ/ (mean = 0.10, SE = 0.004), showed numerically more M100S than the more deviant condition, /i/ (mean = 0.11, SE = 0.004) and /a/ (mean = 0.10, SE = 0.004). However, these comparisons were not significant (/ɪ/ vs. /i/: t(30) = 0.94, n.s.; /ɛ/ vs. /a/: $t(30) = 0.44$, *n.s.*). The only comparison that came out significant showed more M100S for /ɪ/ compared to /e/ ($t(30) = 2.88$, $p = 0.007$), which didn’t correspond to our predictions.

To investigate these comparisons further, the ERF for pairs of vowels were compared separately for each task condition. Figure 8.2b shows which comparisons came out as significant (uncorrected; speaking: /i/ vs. /a/, /ɪ/ vs. /e/, /ɪ/ vs. /a/ and /ɛ/ vs. /a/. The contrasts /i/ vs. /ɪ/ and /ɪ/ vs. /ɛ/ were close to significant ($p = .055$ and $p = 0.087$); listening: /a/ vs. every other condition). Corresponding to our expectations, the production ERF in a more deviant condition like /a/ was higher than the ERF for a less deviant condition like /ɛ/, which could correspond to the numerical difference found in the amount of SIS reported above. Note that when the results were corrected for multiple comparisons (Bonferroni), for production only the /ɪ/ vs. /a/ and the /ɪ/ vs. /e/ comparisons survive (p = .0001 and p = .003). For listening, only the /e/ vs. /a/ comparison survives multiple comparison correction (p = 0.0004).

The results in Figure 8.2b do not fully correspond to our expectations. One unexpected result is that the ERF for /a/ in the listening condition was larger than the ERF for the other vowels (although mean intensity was equalized, peak amplitude of /a/ was significantly larger than peak amplitude of any other vowel, all t(31)'s > 2.79, p < 0.01), but maybe even more surprising is the fact that the amount of M100S was higher for /ɪ/ than for /e/. Given that the stimulus in the /e/ condition should be acoustically closest to what participants started with on every trial, this condition should elicit the most suppression and thus the lowest ERF value in the production task. This was not observed, and is difficult to explain.

A number of post-hoc explanations could perhaps explain this result. First, due to the explicit imitation instructions, people might simply not have expected to hear /e/, because they expected to have to imitate. Second, given that the stimuli were pre-recorded, trial-by-trial variability in the people's initial vowel will have affected the degree of mismatch and therefore the suppression associated with it. Third, if people in our experiment were comparing their motor-based predictions with their actual auditory feedback (rather than the stimuli), neural suppression would be expected not to vary between stimulus conditions. This effect might have been dominant in the M100 modulation, which means the neural suppression would have cancelled out (part of) the error-related neural enhancement. Finally, it is important to note a limitation of our paradigm with respect to this research question. It is possible that during the time window of our analysis, our participants had already issued motor commands to change their ongoing vowel production, so their internal target was not /e/ anymore. However, similar analyses in an earlier (50-150ms) time window revealed a similar pattern of results. Obviously, these are all merely post-hoc explanations and require further investigation.

8.3.3 SIS predicts participants' imitation performance

Despite the fact that M100S did not scale according to the deviance from people's initial speech production, we can still address whether modulation of the M100S is predictive of how much people adapt their speech motor output. To explore this issue, we investigated whether the average amount of M100S across subjects is correlated with subjects' performance on the imitation task (Figure 8.3). We hypothesized that people would vary in their sensitivity to errors, and that this variation would be reflected in the average amount of suppression they show. Less suppression on average should then relate to a higher sensitivity to errors (or being less lenient to prediction errors), which would lead to stronger and/or quicker adjustment.

Corresponding to this prediction, we found that M100S correlated with imitation performance ($r(29) = -0.48$; $p < 0.01$). However, note that participants listened to different stimuli (their own voice), so it would be possible that participants who show less M100S simply heard stimuli that were acoustically further apart. A control analysis showed that the correlation reported above was not due to subject-specific dispersion of the stimuli in vowel space. Specifically, a multiple regression analysis including imitation performance and stimulus dispersion as separate predictors showed a main effect of imitation performance ($t(28) = -2.99$, $p < 0.01$), but not of stimulus dispersion ($t(28) = 0.58$, *n.s.*).

Of note in the current results, seven subjects show a mean negative imitation performance. This effect may reflect non-linearities in vowel space, as subjects could have ended up farther away in acoustic space while still staying within the same perceptual category. Importantly, when these subjects were excluded, we still found a negative trend ($r(22) = -0.39$, $p < 0.06$). Also, note that an important limitation of our study is that we are unable to determine whether less M100 suppression actually reflects less neural suppression, rather than an enhancement of neural activity. So although we found a reliable correlation between M100S and behavioral performance, the neural mechanism driving this correlation remains unclear.

8.4 DISCUSSION

In the current study we investigated M100 modulation in the context of a speech imitation task. Based on current theories of speech-motor control (Guenther et al., 2006; Hickok, 2012; Houde & Nagarajan, 2011), we predicted that the magnitude of M100S would decrease with the acoustic distance from the starting vowel of people's speech production, and furthermore, that M100S would predict how much people adapted their speech in the imitation task. Acoustic measures of the people's speech output during the task showed that subjects complied with the task instructions by adapting their speech to the auditory stimuli. In the MEG data, we found a strong suppression of the auditory M100 in the speaking task relative to the listening task, which replicates SIS in this new paradigm. A closer look at the ERFs in the different conditions revealed that the varying auditory stimuli modulated the amount of suppression, although the results did not correspond to our predictions.

With regards to the second prediction, an analysis of individual differences in the speech imitation task confirmed a relationship between M100S and

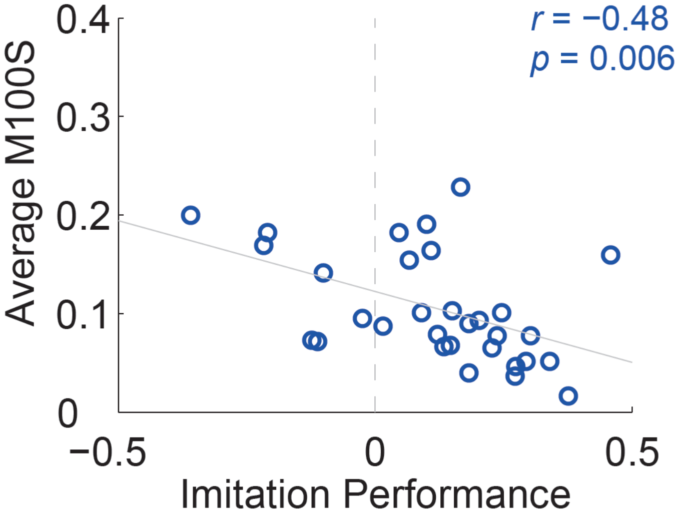

Fig. 8.3 Correlation between the average imitation performance and the overall M100S per subject. Imitation performance is expressed as the proportional change in distance to the target from the start to the end of a trial ((distance_to_targetbegin – distance_to_targetend) / distance_to_targetbegin). The solid grey line represents a linear fit.

participants' performance in the imitation task. Specifically, people with reduced M100S were better at speech imitation. This result is consistent with the models of speech motor control described in the introduction. Here we take the average amount of people's M100S as indicative of their sensitivity to errors. Specifically, some people might be more sensitive (or less lenient) to small errors between their intended speech and acoustic mismatches. As a result, these participants would show on average less M100S. A higher sensitivity to mismatches would lead people to generate error signals more readily, thus driving speech motor adaptation, resulting in better imitation performance.

The findings from the current study are the first to explicitly link individual variation in M100 modulation to speech imitation performance. This result provides an important test of one of the main hypotheses stemming from current theories of speech motor control, and is consistent with M100S indexing an error signal that is used to adapt speech motor behavior. Despite this result, we did not find strong evidence in favor of another prediction stemming from these accounts, namely, that the amount of M100S should decrease with the distance from the original target utterance. As noted above, this may be a result of the current task conditions, and thus remains an open question for future investigation. Nevertheless, the current study highlights the importance of taking individual differences into account when studying speech production and opens up possibilities for further investigations in how inter-individual variation in speech production relates to variation in the psychological and neural mechanisms underlying speech motor control.

REFERENCES

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America, 116*(3), 1729–1738. doi:Doi 10.1121/1.1779271

Bastiaansen, M. C. M., & Knosche, T. R. (2000). Tangential derivative mapping of axial MEG applied to event-related desynchronization research. *Clinical Neurophysiology, 111*(7), 1300–1305. doi:Doi 10.1016/S1388-2457(00)00272-8

Behroozmand, R., & Larson, C. (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. *BMC Neuroscience, 12*(1), 54. Retrieved from http://www.

biomedcentral.com/1471-2202/12/54

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences of the United States of America, 110*(7), 2653–2658. doi:DOI 10.1073/pnas.1216827110/-/DCSupplemental

Flinker, A., Chang, E. F., Kirsch, H. E., Barbaro, N. M., Crone, N. E., & Knight, R. T. (2010). Single-Trial Speech Suppression of Auditory Cortex Activity in Humans. *Journal of Neuroscience, 30*(49), 16643–16650. doi:Doi 10.1523/Jneurosci.1809-10.2010

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301. doi:10.1016/j.bandl.2005.06.001

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport, 17*(13), 1375–1379. doi:10.1097/01.wnr.0000233102.43526.e9

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135–145. doi:Doi 10.1038/Nrn2158

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*(28). doi:10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience, 14*(8), 1125–1138. doi:10.1162/089892902760807140

Kent, R. D. (1974). Auditory-Motor Formant Tracking: A Study of Speech Imitation. *J Speech Hear Res, 17*(2), 203–222. Retrieved from http://jslhr.asha.org/cgi/content/abstract/17/2/203

Liu, H., Meshman, M., Behroozmand, R., & Larson, C. R. (2011). Differential effects of perturbation direction and magnitude on the neural processing of voice pitch feedback. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 122*(5), 951–7. doi:10.1016/j.clinph.2010.08.010

Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences, 94*(20), 10979–10984. doi:10.1073/pnas.94.20.10979

Naatanen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound - a Review and an Analysis of the Component Structure. *Psychophysiology, 24*(4), 375–425. doi:DOI 10.1111/j.1469-8986.1987.tb00311.x

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What Does Motor Efference Copy Represent? Evidence from Speech Production. *Journal of Neuroscience, 33*(41), 16110–16116. doi:Doi 10.1523/Jneurosci.2137-13.2013

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). Fieldtrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience, 2011*(2011). doi:doi:10.1155/2011/156869

Sitek, K. R., Mathalon, D. H., Roach, B. J., Houde, J. F., Niziolek, C. A., & Ford, J. M. (2013). Auditory Cortex Processes Variation in Our Own Speech. *Plos One, 8*(12), e82925. doi:10.1371/journal.pone.0082925

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America, 8*(3), 185–190. doi:10.1121/1.1915893

Stolk, A., Todorovic, A., Schoffelen, J. M., & Oostenveld, R. (2013). Online and offline tools for head movement compensation in MEG. *NeuroImage, 68*, 39–48. doi:10.1016/j.neuroimage.2012.11.047

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology,* 1, 166. doi:10.3389/fpsyg.2010.00166

Ventura, M., Nagarajan, S., & Houde, J. (2009). Speech target modulates speaking induced suppression in auditory cortex. *BMC Neuroscience, 10*(1), 58. Retrieved from http://www.biomedcentral.com/1471-2202/10/58

Vigario, R., Sarela, J., Jousmaki, V., Hamalainen, M., & Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering, 47*(5), 589–593. doi:Doi 10.1109/10.841330

Zwicker, E. (1961). Subdivision of Audible Frequency Range into Critical Bands (Frequenzgruppen). *Journal of the Acoustical Society of America, 33*(2), 248–&. doi:Doi 10.1121/1.1908630

9

GENERAL DISCUSSION

9.1 SUMMARY OF FINDINGS

The current thesis reports studies investigating various aspects of perceptual influences on speech motor control, with a particular emphasis on auditory feedback processing during speech production. In **Chapter 2**, it was shown that across individuals, better auditory acuity is associated with more precise speech production. This suggests that better auditory acuity leads speakers to be stricter in rejecting outlying speech productions, which over time results in more precise articulations (i.e., less within-phoneme variability and higher between-phoneme distances).

Chapters 3 through 7 make use of altered auditory feedback to investigate how speakers' production is affected by perturbed auditory feedback. In general, speakers respond to feedback perturbations in two ways. First, speakers issue immediate, additional motor commands in response to unexpected altered feedback. Second, speakers may adapt feedforward speech motor programs to avoid errors in the future. **Chapter 3** compares responses to altered formant feedback in a consistent and an inconsistent condition to examine whether feedback consistency affects feedback-related responses. It was shown that this is indeed the case, with speakers showing a stronger response when the feedback perturbation was consistent. **Chapter 4** uses a pitch perturbation paradigm to investigate why speakers sometimes follow and sometimes oppose the feedback perturbation direction. It was shown that pitch fluctuations right before a pitch perturbation kicks in affect how speakers respond to the perturbation. This result suggests that the production system may respond to unexpected feedback perturbation by – at least initially – simply doing the opposite of what it was doing at the onset of the perturbation. **Chapter 5** focuses on the neural correlates of unexpected feedback processing during speech production. The results show that an unexpected auditory feedback perturbation leads to increased activity in both auditory and motor-related cortical areas. Increased θ and β band activity over motor-related areas suggest involvement of those areas in auditory-motor integration.

In **Chapter 7**, an altered auditory feedback paradigm was used to examine yet another aspect of the perception-production relationship in speech. Specifically, this chapter investigates whether perceptual learning affects subsequent production learning. As a prequel to this chapter, **chapter 6** investigates whether listeners use visual (lip-read) information to recalibrate vowel categories (audiovisual recalibration). Subsequently, the perceptual learning paradigm used in chapter 6 was used in chapter 7 to investigate the effects on subsequent speech production learning. The results suggest that audiovisual perceptual recalibration does not directly affect speech adaptation in an altered auditory feedback paradigm. The perceptual task did affect speech adaptation in an indirect way, presumably via a process of selective adaptation. The results in this chapter indicate that audiovisual recalibration of phoneme categories may occur at a context-specific level of abstraction, rather than through adaptation of abstract, modality-independent phoneme categories.

While chapters 3 through 7 have focused on auditory feedback processing during speech production, the last experimental chapter, **chapter 8**, considers another case where speech perception influences speech production: speech imitation. Clearly, imitation is a speech act that crucially relies on perceptual input as the perceptual system provides the imitation target. In this chapter, it was examined how the distance between speech output in an imitation task and the imitation target may affect auditory cortical activity. It is shown that auditory processing in the cerebral cortex is different during speaking compared to passive listening, and that the way auditory neural responses are modulated is associated with speakers' performance in a speech imitation task.

9.2 AUDITORY FEEDBACK PROCESSING

Most experimental studies in the current thesis investigated aspects of auditory feedback monitoring during speech production. Importantly, all perturbations of auditory feedback employed in these studies were small, and speakers were not consciously aware of any perturbations. Although current theoretical models of speech motor control are able to account for the main findings from studies using altered auditory feedback paradigms, the studies reported here show that dominant theoretical frameworks (e.g., SFC, Houde & Nagarajan, 2011, DIVA, Tourville & Guenther, 2011) are not able to account for all the data and thus should be adjusted. The research reported in the current thesis highlighted two main issues. The first issue is about the distinction and relationship between short-term responses to unexpected feedback perturbations and longer-term adaptations of internal forward models, which was explored in chapter 3. The second one is concerned with the distinction between following and opposing responses to altered auditory feedback as explored in chapter 4 and 5. Both issues show that the dominant models are not fully able to account for the data.

With respect to forward model adaptation, we suggested that whether or not a mismatch between predicted and observed auditory feedback leads to forward model adaptation depends on the consistency of the auditory feedback: if the prediction error is consistent across trials, participants adapted the forward model, while they did not or less so when the prediction error was inconsistent. Although current models may be adjusted to account for these data, both DIVA and the SFC model are unable to account fully for the data. The DIVA model has no clear way of keeping track of feedback consistency or allowing it to determine speech adaptation. Although adaptation occurs through applying weighted versions of short-term compensatory responses to the feedforward control system, the model is not clear on how these weights are determined or adjusted. In contrast, the SFC model does keep track of the variance of the feedback signal (although variance and consistency are not the same thing; see Gonzalez Castro, Hadjiosif, Hemphill, & Smith, 2014 for a dissociation between them). However, the SFC model is not clear on a distinction between short-term compensatory responses and longer-term adaptation of the forward model, nor is it clear on how it is determined whether longer-term adjustments are made. In addition, while the DIVA model proposes a crucial link between short- and longer-term adjustments (the latter being based on the former), the SFC model allows for separate mechanisms. The lack of a correlation between compensation and adaptation in chapter 3 is more in line with the SFC model in this regard, although this null result should be interpreted with care. Overall, the current results suggest that contemporary models should be adjusted in such a way that the system (1) keeps track of the consistency of feedback information and (2) uses feedback consistency to determine whether the feedforward system should be adapted. This adaptation mechanism should be separate from a mechanism that handles online feedback-based error correction.

Another issue for both DIVA and SFC is their lack of explanation for so-called following responses. Although this type of following response has been reported multiple times in the literature, the dominant models of speech motor control cannot account for following responses. In fact, surprisingly few studies have directly examined what factors play a role in whether speakers follow or oppose an unexpected auditory feedback perturbation. Hain et al. (2000) proposed the distinction has to do with whether speakers consider the perturbation to be self-generated, or originating from an external source, while Schuerman et al. (2017) briefly speculate that the distribution in vowel space may play a role. Specifically, phonemes located at the edge of vowel space (like [i] in their case) may be more likely to lead to following responses, due to increased reliance on somatosensory feedback.

The study reported in chapter 4 explicitly examines the distinction between opposing and following responses. Interestingly, it was shown that most if not all participants made following responses in at least some trials, and that the distribution of following vs. opposing responses varied greatly between participants. This suggests that the type of response (following or opposing) is not solely determined by the putative source of the auditory signal, as there is no clear reason why a listener would consider an auditory signal as externally generated on some trial but as self-generated on the next trial under exactly the same circumstances (cf. Hain et al., 2000). In addition, the main result of the study suggested the response type was associated with the pitch contour right before the pitch perturbation kicked in, showing the speech motor system takes its current state into account in determining how to respond to a pitch perturbation. Specifically, this suggests that the system's initial reaction to a pitch perturbation may be to simply do the opposite of what it was doing: if pitch happened to be increasing, a perturbation would lead the speaker to decrease pitch and vice versa. Current models would have to be adapted to allow for such a mechanism. Further studies would need to examine this idea more closely. This is especially relevant for pitch perturbations, as producing speech at a target pitch is known to involve fluctuating around this target pitch, and so pitch is not stable.

Additional data with respect to the distinction between opposing and following responses to pitch perturbation is offered in chapter 5, where an additional MEG analysis showed increased activity in the supplementary motor area (SMA) during opposing responses compared to following responses. It was suggested that SMA is involved in the initiation of compensatory articulatory responses, roughly in line with the role of SMA as hypothesized in the DIVA model. Note, however, that while DIVA proposes that the SMA is involved in the initiation of speech acts, it does not associate SMA with a role in auditory feedback processing per se. Therefore, in order to account for the current data, models would have to incorporate SMA as well in the neural control of online feedback-based behavioral responses. To our knowledge, this is the first study that examines the neural correlates of opposing vs. following responses to altered auditory feedback. Much is still unclear about what factors determine whether a speaker follows or opposes altered auditory feedback and what the supporting neural network is. The current findings on opposing and following responses suggest at any rate that speech motor control is fundamentally dynamic in nature, and can serve as a starting point for further studies.

With respect to the neural correlates of auditory feedback processing, the results reported in this thesis are more or less in line with dominant models of the neural control of speech with regards to the cortical areas involved in speech production and auditory feedback processing. Chapter 5 shows strong auditory cortical responses to the onset and offset of auditory feedback perturbations, but also extends the literature by examining frequency-domain (θ and β band) neural correlates of feedback processing. These responses were localized predominantly in motor-related cortical areas, suggesting their involvement in sensorimotor integration. Note that while θ band power increases related to enhanced pitch processing have been reported previously (Behroozmand, Ibrahim, Korzyukov, Robin, & Larson, 2015), the increased β band power has to our knowledge not been reported in relation to auditory feedback processing during speech production. Overall, the results from this MEG study suggest that early on, unexpected auditory feedback is detected in the auditory cortices, while behavioral responses may be generated by involvement of motor-related areas as indicated by increased power in the β and θ band. This is roughly in line with both the DIVA and SFC models, where auditory feedback is processed in auditory cortices initially, which in case of a prediction error may lead to transformation into behavioral responses in frontal areas (motor and pre-motor cortices).

We speculatively conclude that β may indicate involvement of the motor system, given previous findings that link β power with motor function and sensorimotor processing (Kilavik, Zaepffel, Brovelli, MacKay, & Riehle, 2013). However, a β increase under altered auditory feedback seems in contrast with a recent proposal for the function of β oscillations in both motor and language tasks (Engel & Fries, 2010; Lewis & Bastiaansen, 2015; Lewis, Schoffelen, Schriefers, & Bastiaansen, 2016). The β maintenance hypothesis proposed by these authors suggests that increased β power reflects maintenance of the current cognitive set, in contrast with what one expects under unexpected auditory feedback that leads to a behavioral adjustment of ongoing speech production. Future research should examine more closely what the role of β band activity is with respect to auditory feedback processing in speech motor control specifically, and whether or not it is related to its role in other cognitive tasks.

Chapter 8 also investigated the neural correlates of speech production, in particular in a speech imitation task. Although not strictly speaking auditory feedback, speech imitation has a lot in common with altered auditory feedback. Both are cases of how speech perception may influence subsequent speech production. While altered auditory feedback involves a manipulation of one's own voice in real-time, the imitation task used in chapter 8 had participants imitate a recorded vowel (produced by the same participant in an earlier session). The results suggest that also in terms of neural correlates, there may be commonalities between auditory feedback processing and auditory processing of externally-generated sounds during speech production. A reduction of the M100, the magnetic counterpart of the perhaps better-known N100, was found during speech production compared to a listening task. A reduction of N100 suppression or an increase in N100 has previously been argued to reflect the workings of a forward model in speech production (Curio, Neuloh, Numminen, Jousmaki, & Hari, 2000; Heinks-Maldonado, Nagarajan, & Houde, 2006; Houde, Nagarajan, Sekihara, & Merzenich, 2002). Especially the finding that individual variability in the suppression of the M100 during the imitation task in chapter 8 was related to variability in imitation performance is in line with studies on auditory feedback. It suggests that, as in auditory feedback processing, there is a crucial comparison between an auditory prediction and observed auditory input that takes place in the auditory cortices.

The studies on auditory feedback presented in the current thesis clarify the workings of auditory feedback processing in speech. The findings suggest that speakers constantly monitor their auditory feedback, which allows them to detect and correct errors online, as well as adapt the speech system to avoid similar errors in the future. In order to determine whether the speech system needs to be adapted in this way, speakers keep track of the consistency of feedback information. This is important, as it relates to how auditory feedback processing may be used for speech learning. With respect to online responses to unexpected feedback, speakers may either oppose or follow unexpected feedback. The current research highlights the dynamic nature of the speech system, by showing that feedback processing does not follow strict serial processing stages, but is directly sensitive to the dynamic state of the speech production system. These results can serve as a starting point for future studies on the factors that determine how speakers respond to unexpected auditory feedback.

9.3 WHAT DOES AUDITORY FEEDBACK TEACH US ABOUT THE PERCEPTION-PRODUCTION LINK?

As was briefly discussed in chapter 1, many authors have argued for a close link between speech perception and speech production, up to and including arguments that speech perception crucially involves the production system (Galantucci, Fowler, & Turvey, 2006; Liberman & Mattingly, 1985). The results from our experiments on auditory feedback in speech production, however, argue at least against an immediate, automatic influence of perception on production. Both chapters 3 and 7 suggest that this influence is context-dependent. Chapter 3 shows that contextual consistency affects the amount of influence perception of auditory feedback has on speech production, while chapter 7 shows no strong evidence that audiovisual recalibration directly affects subsequent speech adaptation. This is not to say that there is no perception-production link, as we do show such an association in several other chapters: chapter 2 shows an association between auditory acuity and speech production variability, chapter 5 shows that auditory feedback processing is accompanied by a neural network consisting of both auditory and motor-related cortical areas, and chapter 8 shows that the cortical response to auditory stimulation is associated with performance in an imitation task. These findings are in line with the vast literature showing that speech perception and speech production are linked (Buckingham H. W., Hickok, & Humphries, 2001; Skipper, Devlin, & Lametti, 2017). However, based on the data from chapters 3 and 7, we argue that there are constraints on the perception-production link in speech. It is indeed to be expected that speech perception's influence on speech production is not direct but constrained. For example, it is easy to learn to understand the speech produced by speakers with a foreign accent, but this does not lead us to produce speech with a foreign accent ourselves.

An important constraint when it comes to the effects of auditory feedback on speech production is the context in which auditory feedback was perturbed. The results in chapter 3 suggest the feedback's influence is modulated by the contextual consistency. In other words, speakers' reliance on auditory feedback varies as a function of its reliability. If auditory feedback is consistently off by a similar magnitude, speakers will adapt to avoid future errors, but when auditory feedback is sometimes off, but sometimes on point, speakers may distrust the signal and therefore rely more on their feedforward motor commands. This is in line with previous studies suggesting the gain on auditory feedback information is variable. In other words, the effect that perturbed auditory feedback has on speech production is modulated by other factors, like vocal training (Zarate & Zatorre, 2008), or whether one speaks one's native language or not (Van Borsel, Sunaert, & Engelen, 2005). This is true not only within individuals depending on the context, as in the current data, but also across individuals (Lametti, Nasir, & Ostry, 2012).

Another modulatory constraint on the influence of auditory feedback is the linguistic relevance of the mismatch between expected and observed auditory feedback. In other words, listeners may assign more weight to a feedback perturbation that affects a task-relevant parameter compared to a task-irrelevant parameter (Chen, Liu, Xu, & Larson, 2007; Mitsuya, MacDonald, Purcell, & Munhall, 2011; Niziolek & Guenther, 2013; Xu, Larson, Bauer, & Hain, 2004). Although the current thesis does not report results that directly target whether task-relevance affects speech adaptation, chapter 5 showed that the auditory cortex responds strongly to perturbation onsets and offsets during speech production, but not when participants were just listening. Such a result may be related to task-relevance as the nature of the pitch matching task meant participants were explicitly paying attention to their pitch, while the listening task merely involved passive listening. As such, the pitch perturbations in auditory feedback may have been highly task-relevant. Similarly, findings in chapter 8 showed increased responsiveness of the auditory cortex to the task-relevant vowel stimuli during the imitation task, compared to the same stimuli during passive listening. The results from chapters 5 and 8 are thus consistent with the auditory cortex being especially sensitive to feedback perturbations because they are relevant to the task at hand, even though the speakers were not consciously aware of any perturbation.

Overall, although it is clear that speech perception and speech production are linked, the current thesis shows that this link is not immediate or automatic, but subject to constraints, for example the current context. The current context may indicate how reliable the feedback information is, or how relevant mismatching or unexpected feedback is. These factors modulate the effect that auditory feedback has on speech production.

9.4 INDIVIDUAL VARIABILITY

In recent years, several authors in the cognitive sciences (Henrich, Heine, & Norenzayan, 2010; Kanai & Rees, 2011) have argued that more efforts should be spent investigating individual variability, rather than focusing on the “average” participant. Indeed, by averaging across individuals and studying only the behavior of the average participant, one could end up ignoring a lot of potentially useful information. The research reported in the current thesis has shown that in many aspects, there is remarkable inter-individual variability. The results allow us to suggest some of the factors that may drive individual variability: individual variability in speech production was associated with variability in speech perception (chapter 2), individual variability in amount of audiovisual recalibration was related to the sharpness of the phoneme boundary in question (chapters 6 and 7), and individual variability in M100 suppression was related to performance in a vowel imitation task (chapter 8). In addition, there was substantial individual variability in how speakers respond to altered auditory feedback, as reported in chapters 3 and 4. All participants in the study in chapter 4, for example, showed both opposing and following responses to pitch-shifted auditory feedback, but there was large variability between individuals with respect to the distribution of opposing and following responses. In fact, the research in chapter 4 is a good example of how we can use individual variability as a tool to test and refine existing theories. While the dominant view on speech motor control can aptly account for the group average response to pitch-shifted feedback, our approach of capturing the variability in behavioral responses showed was able to show where current theories were not able to account fully for speakers' behavior. As is evident from the research reported here, more studies are needed to disentangle individual variability with respect to speech production and speech perception, especially studies with much larger samples than these rather small-scale studies.

9.5 THESIS SCOPE AND LIMITATIONS

Although the aim of the current thesis was to investigate various aspects of speech perception-production interactions, with a special focus on the role of auditory feedback processing during speech production, the scope of this thesis has necessarily left out some important aspects of auditory feedback. This is not to say that these aspects are unimportant or that they do not play a role.

First, all studies in the current thesis have focused on one specific function of auditory feedback, that is to maintain, update and correct speech production (Hartsuiker & Kolk, 2001; Levelt, Roelofs, & Meyer, 1999). Roughly, auditory feedback is considered as being mainly used to make sure the speech sounds produced are in fact the speech sounds that were intended to be produced. However, it is well established that sensory feedback is used to dissociate between self-generated sensations and sensations produced by others as well (Eliades & Wang, 2008; Schütz-Bosbach, Mancini, Aglioti, & Haggard, 2006). Some authors have suggested that auditory feedback plays a role in determining sense of agency during speech production (Lind, Hall, Breidegard, Balkenius, & Johansson, 2014a, 2014b).

Second, the current thesis has focused on the role of auditory feedback in speech production in typical adult speakers. However, dysfunction of sensorimotor processing in speech production may lead to a number of well-known clinical correlates, among which developmental stuttering (Max, Guenther, & Gracco, 2004; Tourville, Cai, & Guenther, 2013), conduction aphasia (Hickok, Houde, & Rong, 2011) and auditory hallucinations in schizophrenia (Ford & Mathalon, 2005; Heinks-Maldonado et al., 2007), and others. In addition, some authors have proposed that deficient perception-production interaction in speech may be at the basis of developmental dyslexia, with dyslexics responding differently to altered auditory feedback (van den Bunt et al., 2017).

Third, the studies reported here limited their focus to the production of single sounds, sometimes embedded in a single word. However, speech production of course almost always involves producing connected speech, phrases and sentences, often including a lively interaction with one or more interlocutors. It is a question for further research how the current findings translate to these much more complex situations. There is some previous research on this topic, leading to the development of various models, including the GODIVA model, an extension of the DIVA model for speech sound sequencing (Guenther, 2016), models that connect sound-level speech production and perception to higher linguistic levels (Hickok et al., 2011), and a model of speech perception and production in dialogue that focuses on how forward models may be used on multiple linguistic levels in interactive language processing (Pickering & Garrod, 2014).

Finally, with respect to the neural correlates of speech perception-production interactions, the research reported here has been limited to activity in areas of the cerebral cortex. In part, this is due to the limitations of magnetoencephalography. However, it is well established that subcortical areas, including but not limited to the basal ganglia, the thalamus and the cerebellum, play a crucial role in sensorimotor production. Note that the DIVA model indeed includes these regions as part of the speech production network (Guenther, 2016). We refer the reader to the broad literature on these topics (Ackermann, 2008; Argyropoulos, 2016; Barbas, García-Cabezas, & Zikopoulos, 2013; Wildgruber, Ackermann, & Grodd, 2001).

9.6 CONCLUDING REMARKS

The research reported in this thesis has attempted to investigate various aspects of speech perception, speech production, and especially their complex and interesting interaction. A particularly interesting case concerns auditory feedback, as this relates to the fact that speech production almost always is accompanied by concurrent speech perception, with necessary interaction. Although many theoretical insights with respect to motor control have been based on the study of fairly simple motor tasks, such as reaching or making saccades, studying a very complex motor system like speech production puts many of these ideas to the test.

Although current dominant views on speech motor control have been well developed over the last couple of decades, the current thesis suggests that models would need to be adapted to fully account for these data. The main issues that need to be addressed involve the effects of environmental consistency and of the current state of the speech system on how auditory feedback affects speech production. In addition, the evidence suggested that there are important constraints on how speech perception might affect speech production. These constraints include the reliability and the relevance of the perceptual information. Furthermore, the data add to current views on how speech motor control is implemented at the neural level. Auditory predictions or targets are compared to incoming auditory signals in the auditory cortices, and in the case of prediction errors, may be transformed into motor responses in motor and pre-motor cortices.

In many cases, in addition to answering our research questions, the data reported here have generated additional questions, which require further study. Therefore, this thesis may serve as a starting point for further investigations into what drives sensorimotor learning, what factors drive following vs. opposing responses to auditory feedback and especially the associated neural underpinnings, what role different neural correlates play in speech production (e.g., β event-related power increases), and whether and how perceptual learning may affect speech production learning. In addition, many aspects of speech perception-production interactions have been left outside the scope of this thesis, yet the current findings may serve as a starting point for further investigations on these issues as well.

What is clear is that speech perception and production interact in complex ways. One of the major examples of this interaction is the fact that speakers, even without conscious awareness, constantly monitor their own speech production. Speech production is undoubtedly influenced by many factors, but our perception of it is clearly an important one. This self-monitoring process requires the involvement of a broad neural network, and includes integration between, amongst others, auditory and motor-related information streams. What is most intriguing about speech may be not just the complexity of these processes, but the speed and effortlessness with which we perform them on a daily basis.

REFERENCES

Ackermann, H. (2008). Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. *Trends in Neurosciences, 31*(6), 265–72. doi:10.1016/j.tins.2008.02.011

Argyropoulos, G. P. D. (2016). The cerebellum, internal models and prediction in “non-motor” aspects of language: A critical review. *Brain and Language, 161*, 4–17. doi:10.1016/j.bandl.2015.08.003

Barbas, H., García-Cabezas, M. Á., & Zikopoulos, B. (2013). Frontal-thalamic circuits associated with language. *Brain and Language, 126*(1), 49–61. doi:10.1016/j.bandl.2012.10.001

Behroozmand, R., Ibrahim, N., Korzyukov, O., Robin, D. A., & Larson, C. R. (2015). Functional role of delta and theta band oscillations for auditory feedback processing during vocal pitch motor control. *Frontiers in Neuroscience, 9*, 109. doi:10.3389/fnins.2015.00109

Buckingham H. W., J., Hickok, G., & Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science, 25*(5), 663–678. doi:10.1016/S0364-0213(01)00048-9

Chen, S. H., Liu, H., Xu, Y., & Larson, C. R. (2007). Voice F[sub 0] responses to pitch-shifted voice feedback during English speech. *The Journal of the Acoustical Society of America, 121*(2), 1157. doi:10.1121/1.2404624

Curio, G., Neuloh, G., Numminen, J., Jousmaki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Human Brain Mapping, 9*(4), 183–191. doi:Doi 10.1002/(Sici)1097-0193(200004)9:4<183::Aid-Hbm1>3.0.Co;2-Z

Eliades, S. J., & Wang, X. Q. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature, 453*(7198), 1102–U8. doi:Doi 10.1038/Nature06910

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology, 20*(2), 156–165. doi:10.1016/j.conb.2010.02.015

Ford, J. M., & Mathalon, D. H. (2005). Corollary discharge dysfunction in schizophrenia: can it explain auditory hallucinations? *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology, 58*(2-3), 179–89. doi:10.1016/j.ijpsycho.2005.01.014

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), 361–377. doi:10.3758/BF03193857

Gonzalez Castro, L. N., Hadjiosif, A. M., Hemphill, M. A., & Smith, M. A. (2014). Environmental consistency determines the rate of motor adaptation. *Current Biology : CB, 24*(10), 1050–61. doi:10.1016/j.cub.2014.03.049

Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA: The MIT Press.

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research, 130*(2), 133–141. doi:10.1007/s002219900237

Hartsuiker, R. J., & Kolk, H. H. J. (2001). Error Monitoring in Speech Production: A Computational Test of the Perceptual Loop Theory. *Cognitive Psychology, 42*(2), 113–157. doi:10.1006/cogp.2000.0744

Heinks-Maldonado, T. H., Mathalon, D. H., Houde, J. F., Gray, M., Faustman, W. O., & Ford, J. M. (2007). Relationship of Imprecise Corollary Discharge in Schizophrenia to Auditory Hallucinations. *Archives of General Psychiatry, 64*(3), 286. doi:10.1001/archpsyc.64.3.286

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport, 17*(13), 1375–1379. doi:10.1097/01.wnr.0000233102.43526.e9

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2010), 61–135. doi:10.1017/S0140525X0999152X

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron, 69*(3), 407–422. doi:10.1016/j.neuron.2011.01.019

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5(28). doi:10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience, 14*(8), 1125–1138. doi:10.1162/089892902760807140

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nat Rev Neurosci, 12*(4), 231–242. doi:10.1038/nrn3000

Kilavik, B. E., Zaepffel, M., Brovelli, A., MacKay, W. A., & Riehle, A. (2013). The ups and downs of beta oscillations in sensorimotor cortex. *Experimental Neurology, 245*, 15–26. doi:10.1016/j.expneurol.2012.09.014

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback. *Journal of Neuroscience, 32*(27), 9351–9358. doi:10.1523/JNEUROSCI.0404-12.2012

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(01), 1–75. doi:10.1017/S0140525X99001776

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex, 68*, 155–168. doi:10.1016/j.cortex.2015.02.014

Lewis, A. G., Schoffelen, J.-M., Schriefers, H., & Bastiaansen, M. (2016). A Predictive Coding Perspective on Beta Oscillations during Sentence-Level Language Comprehension. *Frontiers in Human Neuroscience, 10*. doi:10.3389/fnhum.2016.00085

Liberman, A. M., & Mattingly, I. G. (1985). The Motor Theory of Speech-Perception Revised. *Cognition, 21*(1), 1–36. doi:Doi 10.1016/0010-0277(85)90021-6

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014a). Auditory feedback of one's own voice is used for high-level semantic monitoring: the "self-comprehension" hypothesis. *Frontiers in Human Neuroscience*, 8. doi:10.3389/fnhum.2014.00166

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014b). Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say. *Psychological Science, 25*(6), 1198–1205. doi:10.1177/0956797614529797

Max, L., Guenther, F., & Gracco, V. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary Issues in Communication Science and Disorders, 31*, 105–122. Retrieved from http://www.keck.ucsf.edu/~houde/sensorimotor_jc/LMax04a.pdf

Mitsuya, T., MacDonald, E. N., Purcell, D. W., & Munhall, K. G. (2011). A cross-language study of compensation in response to real-time formant perturbation. *Journal of the Acoustical Society of America, 130*(5), 2978–2986. doi:Doi 10.1121/1.3643826

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience, 33*(29), 12090–12098. doi:Doi 10.1523/Jneurosci.1008-13.2013

Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience*, 8. doi:Artn 132 Doi 10.3389/Fnhum.2014.00132

Schuerman, W. L., Nagarajan, S., McQueen, J. M., & Houde, J. (2017). Sensorimotor adaptation affects perceptual compensation for coarticulation. *The Journal of the Acoustical Society of America, 141*(4), 2693–2704. doi:10.1121/1.4979791

Schütz-Bosbach, S., Mancini, B., Aglioti, S. M., & Haggard, P. (2006). Self and Other in the Human Motor System. *Current Biology, 16*(18), 1830–1834. doi:10.1016/j.cub.2006.07.048

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language, 164*, 77–105. doi:10.1016/j.bandl.2016.10.004

Tourville, J. A., Cai, S., & Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech (pp. 060180–060180). doi:10.1121/1.4800684

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981. doi:10.1080/01690960903498424

Van Borsel, J., Sunaert, R., & Engelen, S. (2005). Speech disruption under delayed auditory feedback in multilingual speakers. *Journal of Fluency Disorders, 30*(3), 201–217. doi:10.1016/j.jfludis.2005.05.001

van den Bunt, M. R., Groen, M. A., Ito, T., Francisco, A. A., Gracco, V. L., Pugh, K. R., & Verhoeven, L. (2017). Increased Response to Altered Auditory Feedback in Dyslexia: A Weaker Sensorimotor Magnet Implied in the Phonological Deficit. *Journal of Speech, Language & Hearing Research, 60*(March), 654–667. doi:10.1044/2016_JSLHR-L-16-0201

Wildgruber, D., Ackermann, H., & Grodd, W. (2001). Differential Contributions of Motor Cortex, Basal Ganglia, and Cerebellum to Speech Motor Control: Effects of Syllable Repetition Rate Evaluated by fMRI. *NeuroImage, 13*(1), 101–109. doi:10.1006/nimg.2000.0672

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America, 116*(2), 1168. doi:10.1121/1.1763952

Zarate, J. M., & Zatorre, R. J. (2008). Experience-dependent neural substrates involved in vocal pitch regulation during singing. *NeuroImage, 40*(4), 1871–87. doi:10.1016/j.neuroimage.2008.01.026

ACKNOWLEDGEMENTS

This thesis is not only the result of my own work. I am indebted to many who have contributed to this endeavor in some way.

First of all, I owe a great deal of gratitude to all my supervisors – Dan, Jan-Mathijs, Frank, James and Peter –, without whom I would not be where I am today. Dan, it is an understatement to say that none of this would have been possible without you. I still remember our first meeting, to discuss the possibility of a master's internship under your supervision, and especially the fact that I came back from that meeting even more enthusiastic than I was going in. Over the months and years that followed, our meetings kept having that effect on me, even when I was yet again frustrated over some data or experiment not working out the way we hoped it would. I will definitely remember our many inspiring meetings, which often lasted longer than planned. Not only have I learned a lot from you, I also enjoyed talking to you both about research as well as about other stuff, hiking after the conference in Geneva, or enjoying Kölsch during a random conference trip to Cologne. This really is our rather than my project, and I am grateful for all your support and mentorship, even after your move across the pond.

Jan-Mathijs, zonder jou had ik ongetwijfeld nog veel vaker vastgezeten in complexe en minder complexe analyses. Je liet me door de bomen het bos zien, ook al leidde dat vaak tot veel langere meetings dan je gepland had. Ik heb erg veel van je geleerd, vooral als het over programmeren, analyses of methodologie gaat. Daarnaast was het ook gewoon plezant samenwerken. Het is sowieso fijn om 's morgens op kantoor te komen met klassieke muziek op de achtergrond ;-).

Frank, it should be clear to anyone that the only reason you're not mentioned as a copromotor on this thesis is because of the university's rule of no more than four (co)-promotors, because you so deserve it for all the time and effort you invested. It is amazing how you managed to jump on board of an ongoing research project and still give me all the support and advice that I needed. I'm also especially grateful for your continuing support, even after you left academia.

James, I want to thank you for welcoming me in your group. I enjoyed and valued your support. You somehow seemed to be able to make time for me in your otherwise busy schedule whenever I needed it, and always gave me useful feedback on my experimental issues, manuscripts or other struggles. I especially appreciate you making sure I didn't lose too much overview of the project, as well as your support in the search for a new position after the PhD.

Peter, bedankt om me de mogelijkheid te geven eerst een masterstage en vervolgens een onderzoeksproject van vier jaar in je groep te kunnen uitvoeren. Je feedback is steeds waardevol en relevant. Het is bijzonder hoe je telkens met slechts enkele bewoordingen de cruciale punten in een presentatie, manuscript of uiteenzetting weet te duiden.

Completing a four-year research project would have been difficult if not impossible without what I had learned beforehand. So I am grateful to all my former teachers and supervisors, especially during my bachelor's in Leiden and my master's in Nijmegen. In het bijzonder wil ik ook Jean-Christophe Verstraete bedanken, die me pakweg 10 jaar geleden aanraadde om in Nederland te gaan studeren, en zonder wie ik waarschijnlijk nooit in Nijmegen was terechtgekomen.

Anyone doing a PhD research project in Nijmegen should count themselves lucky, especially due to the amazing structural support and expertise that is available. I want to use this space for a special shout-out to all people at the Donders Institute, the MPI and the Radboud University who keep these institutes running. Too many people to name, but I definitely want to mention Tildie, Ina, Carolin, Sandra, Nicole, Ayse, Arthur, Berend, thanks for being there to help me with any administrative support and keeping the institute running. Marek, Mike, Uriel, Sander, Jessica, Paul, Erik, Reiner, Alex, Johan, and many others in the technical support groups of the DCCN and the MPI, no doubt many of my experiments would not have been possible without your technical support. Especially thanks to Uriel for helping me out whenever I had technical requests beyond the currently available set-up in the labs.

I am indebted to all my colleagues in the Neurobiology of Language department, for all the fruitful meetings and discussions, feedback and nice social events. I have learned so much from so many of you. Thanks to all my fellow NBL PhDs over the years for their feedback and support, especially to Richard, Gwilym and Dan for trying to keep organizing and/or reviving the PhD meetings, despite everything ;-). I also enjoyed our weekly DCCN-NBL lunches. Thanks, Basil and Joost, for taking the initiative so often, and all the others that regularly joined them, Anne, Matthias, Atsuko, Sophie, Kristijan, and Bohan, among others. Thanks as well to all my colleagues in the Sound Learning department, I've always enjoyed our fruitful meetings and social gatherings. Mark and Will, my fellow speech adapters, I am grateful for sharing your thoughts and issues on auditory feedback. Mark, thanks for joining me to tell people about auditory feedback at the Drongo language festival. Will, thanks for inviting me to collaborate on our MPI innovation grant project together. Thanks to Jana for collaborating on our ASR system, we managed it in the end ;-).

Besides spending my time on research, I organized and taught in 7 "Programming in Presentation" workshops. However, none of the workshops would have been possible without the support of the Presentation support group, thanks to Pascal, Frauke, Ronald, Wilbert, Maarten and Erik.

During my six years in Nijmegen, I have had the pleasure of meeting lots of amazing friends. Thanks to the core members of our more or less regular lunches, including Maarten, Bart, Tobias, Erik and many others. Thanks to Maarten, Erik and Pim for probably way too many and too long *tea* breaks. Thanks as well to everyone who made The Wednesday Quiz such a fun and regular event. Especially all the loyal Fauxkapi members, Erik, Daniel, Peter, Amie, Ashley, Jeff, Lorijn, and many others. Great to keep having these amazing scores (right?), and it was fun being your geography man! Thanks as well to Peter, Amie, Daniel and especially Erik for hosting many great pre-quiz dinners. If only I would have been a better at mariokart or goldeneye :-D. We also had a regular board game group, thanks for all those game nights – some maybe more frustrating than others -, Daniel, Maarten, Pim, Esther and Marleen! I believe it was Maarten who first invited us for a game of Battlestar, of which we won't be speaking anymore for obvious reasons ;-) Looking forward to a game night in Leuven!

In 2012 werd ik lid bij het Nijmeegs studentenorkest, en nog geen half jaar later ging ik mee op tournee naar Muenster en Kopenhagen. Al zijn het er veel te veel om op te noemen, toch in het bijzonder dank aan al de opeenvolgende bestuursleden bij het CMC, al mijn medecelli over de jaren heen - waaronder Joris, Sascha, Franke, Dirk, Anne, Andreas, Tim, Ruth, Tjeerd en Evert-Jan - om de leukste sectie van het orkest te vormen. Dank ook aan mijn medebelgiëciegenoten Olga, Clara, Franka, Martine en Robert, om samen een geweldige mini-tournee naar Leuven en Brussel te organiseren, en vooral ook dank aan iedereen met wie het wel eens gezellig was in de Kluis, waaronder Elise, Loes, Laurens, Maaike, Charlotte, Olga, Joris, Clara, Sascha, Ike, Dirk, Mirjam, Andreas, Tim, Franka, Martine, Brandaan, Martijn, Victor, Emma, Desiree, Simeon, Stijn, Mart, Rikko, Inge en vele anderen. Ook in andere orkesten heb ik vaker meegespeeld, met name in het Nederlands Strijkersgilde en het Nieuw Nijmeegs Kamerorkest. Dank aan alle bestuursleden voor het organiseren van elk project, en dank aan alle vrienden die ik daar heb leren kennen!

Naast deze orkesten speelde ik ook vaak kamermuziek. In het bijzonder wil ik daarvoor Anima, Matthijs en Emiel bedanken om samen het Bowmore kwartet te vormen, en later Anima, Jasmijn, Emiel en Barbara voor het Bowmore kwintet. Het was super om ons eigen concert te organiseren! Verder mogen Franke, Dirk en Joris uiteraard niet ontbreken voor alle hilarische cellokwartetavonden. Ik ga onze repetitieavonden zeker missen!

De resultaten in deze thesis worden voorafgegaan door een prachtige cover. Bedankt, Franke, voor het alweer mooie ontwerp.

Erik and Daniel, thanks for being my paranymphs. We met when I just arrived in Nijmegen, and you've been there all the way, despite the lack of rabbit with plums. I can't believe it's been over six years already! Erik, you truly are the organizer. Thanks for taking initiative for the countless diners, beers, concerts and/or comedy nights, as well as for being your calm, hilarious yet down-to-earth self! Daniel, you've even had the misfortune of being my housemate for a while. Thanks for staying around nonetheless, thanks for joining on all those trips, thanks for enjoying the Dude as much as I do, and thanks for giving the world ABBA. You are one hilariously sarcastic Swede.

Tenslotte wil ik mijn hele familie bedanken voor het vertrouwen en de steun die ik van ze kreeg. Tom, Pieter en Wouter, bij jullie vind ik steeds weer de nodige ontspanning, waar ook ter wereld. Mama en papa, het valt haast niet uit te drukken hoeveel dankbaarheid ik jullie verschuldigd ben. Voor alle steun, al het vertrouwen, zelfs wanneer ik plots besluit om in Nederland te gaan studeren. Bedankt voor alles.

Daphne, schatteke, waar moet ik beginnen? Gij stondt en staat altijd opnieuw klaar met steun, goede raad of een luisterend oor. De afgelopen drie jaar waren ongelofelijk. Bedankt voor het delen van uw taal-nerdiness, uw passie voor muziek, al dat lekker eten en uw zin voor avontuur. Ik zien u graag.

NEDERLANDSE SAMENVATTING

Spraak is een intrigerend fenomeen. We spreken en luisteren elke dag, zonder dat we ons bewust zijn van de complexe processen die in ons brein plaatsvinden. Dit promotieonderzoek richt zich op de interactie tussen spraakproductie en spraakperceptie, en in het bijzonder op de invloed van onze eigen spraakklanken op ons spreken. Wanneer we spreken horen we tegelijkertijd het geluid van onze eigen spraak. Dit noemen we auditieve feedback. Om de rol van auditieve feedback bij het spreken te onderzoeken, lieten we proefpersonen spreken terwijl we hun auditieve feedback manipuleerden. Zo hoorden proefpersonen zichzelf tijdens het spreken bijvoorbeeld op een andere toonhoogte, of hoorden ze zichzelf net wat anders zeggen. Vervolgens onderzochten we welke invloed deze gemanipuleerde auditieve feedback had op de uitspraak van de proefpersonen.

De resultaten van dit onderzoek tonen aan dat sprekers onbewust voortdurend hun auditieve feedback – hun eigen spraakgeluid – monitoren. Die auditieve feedback wordt gebruikt om eventuele spreekfouten op te sporen en te verbeteren, of om het spraaksysteem aan te passen om dergelijke fouten in de toekomst te vermijden. De resultaten van het onderzoek leiden tot onder meer twee belangrijke bijdragen aan de huidige inzichten in de rol van auditieve feedback. Ten eerste speelt de betrouwbaarheid van auditieve feedback een rol. Als de informatie die in auditieve feedback vervat zit niet erg betrouwbaar is, gebruiken we die feedback ook niet om opgespoorde fouten te verbeteren. Denk bijvoorbeeld aan gesprekssituaties met veel achtergrondlawaai. Ten tweede toont dit onderzoek aan dat spraakproductie en de invloed van auditieve feedback een erg dynamisch proces is. De reactie van een spreker op onverwachte auditieve feedback hangt namelijk af van wat zijn of haar articulatie op dat moment was. Als de auditieve feedback niet correspondeert aan de verwachtingen, is de initiële neiging van sprekers om qua articulatie het tegenovergestelde te doen van wat ze aan het doen waren. Ging hun toonhoogte omhoog, dan is de eerste reactie een lagere toonhoogte, en vice versa. Deze resultaten tonen aan dat huidige modellen niet afdoende zijn om spraakproductie te verklaren, en dus aangepast dienen te worden.

In het algemeen zien we dat hoewel spraakperceptie weliswaar een complexe invloed uitoefent op spraakproductie, deze invloed niet automatisch is, maar onderhevig aan beperkingen. Dit inzicht heeft in de eerste plaats theoretisch belang, met gevolgen voor de bestaande spraakproductiemodellen. Inzichten met betrekking tot het verband tussen spraakperceptie en spraakproductie zijn echter ook relevant op vlak van taalonderwijs en spraaktherapie. Het onderzoek dat in het kader van dit promotieproject gevoerd werd kan als startpunt dienen voor vervolgonderzoek met focus op dergelijke toepassingen.

PUBLICATIONS

Franken, M. K., Hagoort, P., & Acheson, D. J. (2015). Modulations of the auditory M100 in an imitation task. *Brain and Language, 142*, 18-23.

Franken, M. K., McQueen, J. M., Hagoort, P., & Acheson, D. J. (2015). Assessing the link between speech perception and production through individual differences. In *Proceedings of the Congress of Phonetic Sciences*. Glasgow, UK: Glasgow University.

Franken, M. K., Eisner, F., Schoffelen, J.-M., Acheson, D. J., Hagoort, P. & McQueen, J. M. (2017). Audiovisual recalibration of vowel categories. In *Proceedings of InterSpeech*. Stockholm, Sweden.

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *The Journal of the Acoustical Society of America, 142*(4), 2007-2018.

FORTHCOMING

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., & Eisner, F. (submitted). Opposing and following responses in sensorimotor speech adaptation: Why responses go both ways.

Franken, M. K., Eisner, F., Acheson, D. J., McQueen, J. M., Hagoort, P., & Schoffelen, J.-M. (submitted). Self-monitoring in the cerebral cortex: neural responses to pitch-perturbed auditory feedback during speech production.

Franken, M. K., Eisner, F., Acheson, D. J., Hagoort, P., & McQueen, J. M. (in prep.). Feedback consistency determines altered auditory feedback processing.

CURRICULUM VITAE

Matthias Franken (born in Antwerpen, Belgium on 7 September 1989) received his bachelor's degree in Linguistics in 2011 from Leiden University. His bachelor's research project was entitled "When single-word spelling breaks down: a case study of acquired dysgraphia". During his time in Leiden, he took a double minor in Native American Linguistics and in Brain and Cognition. The latter already indicates an interest in cognitive neuroscience, which he pursued further with a Research Master's programme in Cognitive Neuroscience at Radboud University in Nijmegen (2013). His Master's research internship was carried out at the Neurobiology of Language department of the Max Planck Institute for Psycholinguistics, entitled "Modulation of speaking-induced suppression in speech imitation".

DONDERS GRADUATE SCHOOL FOR COGNITIVE NEUROSCIENCE

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit:

http://www.ru.nl/donders/graduate-school/phd/