Perturbed Speech

Jana Brunner

Perturbed speech

How compensation mechanisms can inform us about phonemic targets

Jana Brunner

31 mai 2009

Abstract

The present study describes the results of a 2 week perturbation experiment where speakers' vocal tract shape was modified due to the presence of an artificial palate. The aim of the work is to investigate whether speakers adapt towards acoustic or articulatory targets. Speakers were recorded regularly over the adaptation time via electromagnetic articulography and acoustics. Immediately after perturbation onset speakers' auditory feedback was masked with white noise in order to investigate speakers' compensatory behaviour when auditory feedback was absent.

The results of acoustic measurements show that in vowel production speakers compensate very soon. The compensation in fricatives takes longer and is in some cases not completed within the two weeks. Within a session and for each speaker the sounds can be distinguished solely by acoustic parameters. The difference between the session when no auditory feedback was available and the session when auditory feedback was available was greater for vowels with less palatal contact than for vowels with much palatal contact. In consonant production auditory feedback is primarily used in order to adapt sibilant productions. In general, adaptation tries to keep or enlarge the articulatory and acoustic space between the sounds. Over sessions speakers show motor equivalent strategies (lip protrusion vs. tongue back raising) in the production of /u/. Measurements of tangential jerk suggest that after perturbation onset there is an increase in articulatory effort which is followed by a decrease towards the end of the adaptation time.

The compensatory abilities of speakers when no auditory feedback is available suggest that speakers dispose of an articulatory representation. The fact that motor equivalent strategies are used by the speakers, however, supports acoustic representations of speech. It is therefore concluded that articulatory representations belong to the speech production tasks. However, since they are modified as soon as the acoustic output is not the desired one any more, they rather function in the domain of movement organisation and the acoustic representations dominate.

Keywords:
articulation, perturbation, compensation, movement optimisation, phoneme representations

Zusammenfassung

Die Studie befasst sich mit der Adaption der Artikulation als Folge einer insgesamt zweiwöchigen Veränderung der Vokaltraktgeometrie durch einen künstlichen Gaumen. Ziel der Arbeit ist zu untersuchen, ob die Adaption auf artikulatorische oder akustische Ziele hin erfolgt. Die Produktionen der Sprecher wurden während der Adaptionszeit regelmäßig akustisch und per elektromagnetischer Artikulographie aufgenommen.

Akustische Analysen haben gezeigt, dass die Vokalproduktion sofort nach Perturbationsbeginn adaptiert wird. Für die Adaption der Frikative benötigen die Sprecher mehr Zeit, in einigen Fällen ist die zweiwöchige Adaptionszeit nicht ausreichend. Wenn die Daten nach Sprecher und Aufnahme getrennt betrachtet werden, nehmen die Produktionen einzelner Phoneme abgrenzbare Regionen im akustischen Raum ein. Der Einfluss der auditiven Rückmeldung ist stärker bei Vokalen mit weniger linguo-palatalem Kontakt als bei Vokalen mit viel Kontakt. Bei den Frikativen scheint die auditive Rückmeldung vor allem für die Sibilantenproduktion von Bedeutung zu sein. Generall hat die Adaption zum Ziel, die Abstände zwischen den Lauten beizubehalten oder zu vergrößern. Untersuchungen zur Artikulation des /u/ zeigen, dass die Sprecher über die Sitzungen hinweg motorisch äquivalente Strategien benutzen (Lippenvorstülpung versus Hebung des Zungenrückens). Messungen des Rucks (engl. *jerk*) für artikulatorische Gesten deuten darauf hin, dass der artikulatorische Aufwand nach Perturbationsbeginn steigt und zum Ende der Perturbation hin wieder fällt.

Die Fähigkeit der Sprecher zu kompensieren wenn keine auditive Rückmeldung vorhanden ist, zeigt, dass Sprecher über artikulatorische Repräsentationen verfügen. Die Tatsache, dass motorisch äquivalente Strategien von den Sprechern genutzt werden, unterstützt jedoch akustische Repräsentationen der Phoneme. Die Schlussfolgerung, die aus der Untersuchung gezogen wird, ist daher, dass artikulatorische Repräsentationen beim Sprecher existieren, dass sie aber vor allem der Bewegungsorganisation dienen. Sobald das akustische Resultat nicht mehr das gewünschte ist, beginnen die Sprecher, die Artikulation zu verändern.

Schlagwörter:
Artikulation, Perturbation, Kompensation, Bewegungsoptimierung, Phonemrepräsentationen

Acknowledgements

This book is a slightly modified version of my doctoral thesis which was accepted by the Institut für deutsche Sprache und Linguistik of the Humboldt-Universität zu Berlin and the Institut Nationale Polytechnique de Grenoble under the title *Acoustic compensation and articulo-motor reorganisation in perturbed speech*.

The work was carried out in the framework of a Berlin-Munich co-operation project funded by the Deutsche Forschungsgemeinschft (grants PO 334/4-1 and HO 3271/1-1). Further financial support came from the Ministère délégué à l'enseignement supérieur et à la recherche for a Cotutelle de thèse carried out between the Humboldt-Universität zu Berlin and the Institut Nationale Polytechnique de Grenoble. The recordings were carried out in the phonetics laboratories at the Zentrum für Allgemeine Sprachwissenschaft Berlin and at the Institut für Phonetik und Sprachverarbeitung Munich.

I would like to thank my supervisors Bernd Pompino-Marschall and Pascal Perrier for continuous help over the time of the project, for discussion of preliminary results and for insightful comments on a previous version of this work. Furthermore, many thanks to Jörg Dreyer from the Zentrum für Allgemeine Sprachwissenschaft for carrying out the recordings together with me and for introducing me to EMA preprocessing. Also, I would like to thank Phil Hoole from the Institut für Phonetik und Sprachverarbeitung Munich for providing part of the data discussed here, but also for continous help over the course of three years during which the project was carried out. This involved among other things help with the planning of the experiment, preprocessing of the EMA data, providing segmentation programs and help with the set up of these, discussing preliminary results and commenting on an earlier version of this work.

Furthermore, I would like to thank the student co-workers Olessia Panzyga and Vivien Hein for formant measurements, acoustic and articulatory segmentation. Danke, Roman, für nächtliche Datenrettungsaktionen, and many thanks to Christian Geng, Christian Kroos and Joanna Rycko for all kinds of help with LaTeX problems. Last but not least many thanks to the subjects

who agreed on taking part in a whole series of experiments and on living with a nasty prosthesis in their mouth for two long weeks.

Preface: Aim and structure of the study

This study presents the results of a speech perturbation experiment where the vocal tract shape of seven speakers was modified by a palatal prosthesis and speakers adapted their articulation to this morphological change with the aim to produce speech which sounds normal to them. The adaptation time was two weeks. Speakers wore the palatal prosthesis all day long and practiced speaking. Their articulation was recorded via 2D (five speakers) or 3D (two speakers) electromagnetic articulography. In the initial perturbed session speakers' auditory feedback was masked in order to investigate speakers' compensatory abilities when only tactile feedback was available. The data were analysed in an effort to find out whether speakers adapt towards articulatory or towards acoustic targets. The results of this study therefore contribute to the knowledge of productional targets, but possibly also to the one of perceptual targets. They furthermore provide new insights in speech motor control and movement organisation.

The work was carried out in the framework of the DFG-project *Akustische Kompensation und artikulo-motorische Reorganisation bei künstlich veränderten Gaumenformen*. This project is a co-operation between the Institut für Phonetik und Sprachverarbeitung of the Ludwig-Maximilians-Universität Munich and the Institut für deutsche Sprache und Linguistik of the Humboldt-Universität zu Berlin. Part of the data discussed here were provided by the Munich partner (two speakers), the rest of the data was recorded at the Zentrum für Allgemeine Sprachwissenschaft in Berlin. The primary analysis steps (acoustic segmentation, formant measurements and articulatory segmentation) were carried out to equal parts in Berlin and Munich.

The study deals with articulatory data and discusses the nature of phonemic targets in speech production. The final aim of this work is thus to shed light on the question whether the aim of a speaker is to produce a certain acoustic output or a certain articulatory movement. More concretely, the question is: What is the goal of an adaptation process and what is only an

auxiliary means? Do speakers aim at producing a certain acoustic output and the articulatory movements are the auxiliary means or do speakers aim at producing articulatory movements and the acoustic output is only the consequence of it?

While an answer to this question gives insights into speech production mechanisms it might add knowledge to the perceptual area as well, i.e. to the question whether what is perceived by a listener are the articulatory movements or whether the linguistic message is directly in the acoustic signal. More concretely, the perceptual question to which this work might contribute is whether listeners who hear a sound directly perceive the articulatory action or event which is transmitted by the acoustic signal or whether they take the linguistic message directly from the acoustic signal and the articulatory movements are just a means to produce them. Thus, it is again the question of what is primary and what is an auxiliary means: If one assumes that a listener perceives articulatory movements, these articulatory movements are primary and the acoustic signal is the auxiliary means, it just transmits the signal. If one assumes that the listener takes the message from the acoustic signal, this acoustic signal is primary and the articulatory movements are just the auxiliary means: They are used in order to produce the acoustic signal.

While these questions at first sight seem to be rather theoretical, they bear important implications for a number of other research questions in the field of phonetics and linguistics. To give an example, explanations of language change always presuppose that perceptual and productional phoneme representations are either acoustic or articulatory. Ohala [1996], for example, discusses a case of nasalisation of vowels which are not surrounded by nasals, but by fricatives. He explains the phenomenon by the acoustic consequences of the spread of a glottal opening for the fricative into the vowel which are similar as they would be for a nasalisation. Consequently, if listeners perceive acoustics directly, they perceive the sound as nasalised and start to produce nasalised sounds.

Another example from the opposite direction is the analysis of the "gestural weakening" during the High German consonant shift where stops changed into fricatives (Fowler [1996]). Fowler interprets this shift as support for the perception of articulatory action. This articulatory action is nearly the same for the fricative and the stop whereas the acoustics of the two sounds differs immensely. During the sound shift there must have been a time when both sounds were perceived as "the same", and this is hard to explain if one assumes acoustic representations since the acoustics differs so much for the two sounds.

A further area which would profit from a solution to the question whether

phoneme representations are acoustic or articulatory are all linguistic questions basing their argumentation on variability. Investigating acoustic variability in vowels cross linguistically and drawing conclusions with respect to the vowel systems in the languages of the world, for example, presupposes that what is intended by speakers and perceived by listeners is acoustic in nature. For example, the reason for finding /i, e, a, o, u/ as the most frequent five vowel system rather than /y, ʌ, a, ɣ, ɯ/ has been assumed to be the maximal acoustic distinctiveness of the first five vowels. The second five vowels are acoustically less distinct (Liljencrants and Lindblom [1972]). Articulatorily, the differences between the vowels are the same in both systems.

Investigating the nature of phoneme representations is thus interesting for several reasons, and there is evidence for both assumptions, so that one could say that actually both must exist in the speaker's phoneme representation, articulatory and acoustic components. What should be interesting then is whether one of the two dominates. A way to investigate this is to look under which circumstances speakers vary the articulation. Previous investigations have shown that speakers use motor equivalent strategies, i.e. several articulatory strategies which lead to the same acoustic result. For example, most speakers produce /u/ with two constrictions, a labial and a velar one. When speakers protrude the lips a bit more and at the same time lower the tongue back a little, the acoustic output stays constant. This phenomenon was investigated by e.g. Perkell et al. [1993]. The existence of these motor equivalent strategies suggests that what is perceived by listeners must be acoustic since this is what is constant over different articulatory strategies.

A perturbation experiment seems to be another useful means to investigate this question because the speaker has to relearn articulation. In contrast to normal speech motor learning in children, however, when investigating relearning by adult speakers it is possible to record their normal articulation and acoustics before the relearning starts. Thus, in contrast to first language acquisition, one can record the potential targets (articulatory or acoustic) which the speaker might try to reach. When the speaker then adapts one can investigate whether he adapts towards the acoustic or the articulatory targets, i.e. whether he reaches the initial values in acoustic or articulatory space.

It has to be stressed that this work, even if it might provide support for perceptual questions, centres on speech production. It therefore has to deal not only with the acoustic or articulatory information which the speaker wants to transmit to the listener but it also has to deal with questions of movement organisation and motor control. One of these questions is whether the low articulatory variability which is often found has to do with the fact that speakers select certain optimal strategies and do not use other ones

which would lead to the same acoustic output.

Thus, this work will deal with two different aspects of speakers' phoneme representations. The first aspect is the information transmitted to the listener; the second aspect is the motor control aspect. This leads to a partly dual structure of the study as it is presented here. In the introductory part there is one chapter on perceptual primitives (chapter 1), introducing the main question. This chapter is followed by one about the seemingly unrelated movement optimisation (chapter 2). Afterwards, movement optimisation is not further discussed until chapter 11. The introductory part is rounded off by a third chapter (chapter 3) which discusses previous perturbation experiments. It helps to justify the experimental setup in the present study and to set up expectations for the present work.

After the introductory part, consisting of chapters 1 to 3, the experiment and some basic analysis methods are described in chapter 4. The chapters thereafter each present the results of an analysis bound to a certain question or hypothesis related to the speech production tasks. These chapters are rather independent from each other; they are in general not arranged in a way to complete one another. Chapter 5 investigates the basic requirement for the existence of perceptual primitives in the acoustic domain, namely whether vowels can be distinguished from each other solely by acoustic properties. Furthermore, the development of the acoustic properties over sessions is investigated with the aim to see whether there is a development towards the unperturbed values. Chapter 6 investigates a possible development towards greater articulatory and acoustic distances between the vowels. Chapter 7 investigates the same two questions as chapter 5 but whereas the preceding chapter has dealt with vowels, this chapter is concerned with fricatives. Chapter 8 looks at the influence of auditory feedback on the productions of the vowels. Chapter 9 deals with early adaptation in fricatives, again with special attendance paid to the influence of auditory feedback. The aim of the analyses presented there is to see how much adaptation can be carried out without auditory feedback and for which articulatory tasks auditory feedback is absolutely necessary. Chapter 10 investigates motor equivalent strategies in /u/ in order to see whether speakers, if more than one articulatory strategy could be used, indeed use more than one strategy. Chapter 11 returns to movement optimisation and shows under which circumstances speakers optimise their movements. Finally, chapter 12 summarises the results and draws conclusions.

Contents

Chapter 1

Perceptual primitives: The information transmitted to the listener

Speech is an activity whereby articulators are moved in a certain way and an acoustic signal is produced. The aim of this activity is to convey a linguistic message. Humans have developed a variety of methods to convey messages: speaking, writing, drawing pictures, pointing to something, looking at something, mimicing somebody, to name just a few. Not all these activities might be entirely comparable. For example, understanding speech is a dynamic process whereas the perception of read speech is rather static. Still, all these activities with the aim to transmit messages involve movements of the body, but not all bodily movements are meant to convey meaning. When we write it is not the activity which carries the meaning but rather the result of this activity, but when we mimic someone, it is certainly the activity which carries meaning. For speech things seem less clear. Does the result of the movement, the acoustic signal, contain the message or the movements of the articulators?

While the first assumption might intuitively seem more likely, looking at the acoustic signal as such shows that what we hear seems not to be in there. We have the impression to hear sounds, one after the other, as in a string (Fowler et al. [1980]). The acoustic signal, however, changes continuously. Another problem with assuming that the acoustic signal carries the message is that it differs a lot from speaker to speaker. The signal of a small child is completely different from the signal an adult is able to produce even if we might hear the same message. What the baby and the adult share, however, are similarities in the articulatory movements. So maybe the movement carries the message?

This question about what carries linguistic information and how we reach this information has split researchers into several groups. The most important questions in this debate are: What are the physical correlates of speech percepts? In which domain, acoustic or articulatory, can they be found? How are they decoded? Whereas there are researchers who believe that perceptual primitives are essentially acoustic, others think that what is represented in the brain is motor-driven or articulatory.

One could wonder why researchers care about these questions, but the decision for one or the other direction has important implications. Since the relation between articulation and acoustics is non-linear (Stevens [1972]), it is not possible to, for example, make statements about optimal strategies, accuracy or stability of control without making clear to which domain they are applied (Perrier [2005], p.119f). For instance, if one compares the variability of productions of a phoneme in order to draw conclusions about the influence of the size of the phoneme inventory on the variability one needs to decide for a domain, either the acoustic or the articulatory one. Dixon [1980], for example, looked at Australian languages with a very small vowel inventory and found that there the allophonic variation is huge. He reasoned that this must be because in spite of this variability the perceptual distance between the sounds is still large enough. Tabain and Butcher [1999], however, found for two Australian languages with seven and six places of articulation that the degree of coarticulation, which could be seen as a kind of acoustic variability, is the same as for English, which has only three places of articulation. By looking at acoustic variability these authors imply the assumption of acoustic representations of speech production tasks.

The greatest part of this chapter will be devoted to the description of several theories of speech perception, either stating that perceptual primitives are articulatory (section 1.1), or stating that they are acoustic (section 1.2). Although all these approaches have important implications for other questions as well the focus will be laid on the question of phoneme representations. Supporting experiments for each approach will be described briefly.

In the due course of this study the terms *phoneme representations* and *perceptual primitives* will be used in order to describe roughly what has been termed the *common currency* (introduced among others by Goldstein and Fowler [2003]). To communicate, speakers and listeners share a common currency in the distal physical space, which carries linguistic information. From this perspective, for an efficient speech communication, the task of the speakers is to control their speech production system in order to generate the common currency representing the language units they want to transmit. Listeners, on the other hand, will look in the common currency for the perceptual primitives from which they can recover the intentions of

the speaker. In contrast to that, the term *speech production tasks* will be used in order to describe not only what the speaker wants to convey to the listener, but also the components which are necessary in order to produce articulatory movements (e.g. motor commands). Speech production tasks are consequently the common currency plus motor information.

1.1 The movement carries the message: Perceptual primitives are articulatory

The two most influential theories stating that the objects of speech perception are articulatory are the Motor Theory (Liberman et al. [1967]) and the theory of Direct Realism (Fowler [1986]). Even if there are fundamental differences between the two theories, they share a common main claim, namely that perceptual primitives are in the articulatory domain. The *Motor Theory* states that the perceptual primitives are articulatory gestures, furthermore, that these articulatory gestures are translated by an innate speech module which can find out about the underlying gestures even when these gestures in reality overlap due to coarticulation.

Support for the claim that what we perceive are gestures has been seen in the observation that spectrally equal sounds are heard as different sounds depending on the surroundings. Cooper et al. [1952] describe an experiment where stops with the same burst frequency were presented to listeners. Perception depended on the vowel following the stop. If this sound was a high front or back vowel listeners heard the stop as /p/. If, on the other hand, the vowel was /a/ they heard the stop as /k/. Although Cooper et al. do not go as far yet, one could interpret this finding as supporting the perception of gestures: Even if the acoustic signal is the same the gestures which could have produced it must differ. That is why the stimuli are heard differently.

In the original version of the Motor Theory the phoneme representations are mirrored in invariant measurable characteristics for each gesture, i.e. particular muscle contraction commands. Harris [1974] as well as MacNeilage and deClerk [1969], however, contradicted the existence of these invariants by giving evidence from electromyographic studies[1]. For the Revised Motor Theory (Liberman and Mattingly [1985]) the invariants were therefore the more abstract, intended gestures.

Support for the second claim, that gestures are translated by an innate speech module, comes from duplex perception experiments. In this kind of experiment an isolated synthetic formant transition is presented to one ear

[1] method for recording extracellular field potentials produced by muscles (EMG)

while the other ear receives the "rest" of a CV-stimulus (Liberman [1979]). In a speech stimulus the isolated transition is heard twice: As part of the speech stimulus and as a chirp like sound. If, however, a part of the stimulus is perceived twice, there must be two distinct ways of perceiving the signal, a phonetic and an auditory one.

The second theory which is concerned with articulatory representations, the theory of *Direct Realism*, contradicts this last claim of the Motor Theory, that speech is special and states that, instead of being a special kind of perception, speech perception is rather like any other kind of human perception: We experience the physical cause of a sensation rather than the sensation itself. With regard to speech perception this means that we perceive events which "causally structure air" (Fowler [1996], p.1732).

Support for this theory was given by Fowler and Brown [1997] who looked at intrinsic pitch. Following two observations, namely that intrinsic pitch is higher for high vowels than for low vowels (e.g. Silverman [1987]) and that this difference is not perceived (e.g. Reinholt-Peterson [1986]) as a difference in tone height, Fowler & Brown hypothesized that listeners "parse" f0 along gestural lines. Several production and perception experiments with spoken and sung vowels show that when a high and a low vowel (/i/ and /a/) have the same f0 the high vowel is perceived as lower in frequency. In line with that, speakers in this experiment produced high vowels with a higher f0 than low vowels when they were asked to produce them at the same pitch. However, for spoken vowels, the difference in perception was much smaller than the difference in production, i.e. whereas speakers perceived a vowel pair which differed by about 1 Hz as equal in pitch, they produced a difference of 13-14 Hz when they were asked to produce vowels which match in pitch. In a direct realist perspective the results are interpreted as perception of the cause of a sensation (with the tongue being in a certain position which has, as a secondary effect a certain f0) and not the sensation itself - the intrinsic pitch (see also Fowler [1996] for the interpretation of the result).

Another piece of evidence for gestural representations Fowler and colleagues see in the McGurk-effect. As McGurk and MacDonald [1976] have shown, if speakers are confronted with two contradicting pieces of information, e.g. the acoustic stimulus /ba/ and the visual stimulus of a mouth saying /ga/ they will report having heard /da/. According to researchers claiming that representations are acoustic this effect is the result of the experience listeners have in seeing and hearing people speak at the same time so that the visual and the acoustic cue for a phoneme are associated. Fowler and Dekle [1991], however, offered an alternative interpretation of the effect. In their view it arises because "the optical and acoustic information is convincingly about the same speech event", namely a certain articulatory gesture

(Fowler and Dekle [1991], p.817). Speakers perceive this event and consequently merge the acoustic and the articulatory information in the articulatory domain. In order to experimentally test this hypothesis they set up two cross-modal pairings of stimuli. The first pairing was auditory-orthographic: Subjects were seeing the spelling of a syllable and at the same time they heard the syllable being spoken. This pairing is based on convention and literate subjects have some experience in being offered these two modi at the same time. The second pairing was auditory-tactile: Subjects heard a stimulus and at the same time felt the lips of a model speaker producing a syllable. This pairing of stimuli is, in contrast to the first pairing, not conventional but based on lawful causation, and speakers had no experience in this pairing. It was found that the orthographic stimuli hardly influenced the perception of the auditory stimuli, the tactile component, however, clearly influenced the perception of the auditory component. Fowler & Dekle interpreted this result as supporting the perception of causally related events rather than events which have been related because they have been experienced together very often.

Certain sound change phenomena which offer a problem for theories searching for acoustic representations can easily be explained by Direct Realists. One such phenomenon is the High German consonant shift, where stops changed into fricatives. The reason for this change cannot be perceptual confusion of acoustic characteristics since these vary extremely for stops and fricatives. Treating this change as a weakening of gestures and thus a temporary confusion of two very similar gestures, however, explains the phenomenon (Fowler [1996]).

For Direct Realists the fact that listeners perceive articulatory events (gestures) does not imply that there is no invariance in the acoustic signal because the signal is needed in order to transmit the information about the articulatory actions. According to Fowler, the difference between theories searching for perceptual primitives in articulation and theories searching for them in acoustics is that for the first group the signal is a "specifier of a speech event" whereas for the other it is the "perceptual object itself" (Fowler [1996], p.1773). Direct realists therefore do not see themselves confuted by the evidence provided for acoustic representations. Ohala [1996], for example, discusses the preference of obstruent over sonorant consonants in the languages of the world and states that obstruents are preferred because they are acoustically more salient. Fowler [1996] agrees with him but states that the higher acoustic salience leads to an easier perception of the underlying articulatory action.

A question which arises is *how* people could perceive articulatory gestures directly. A mechanism which could make this possible is the *mirror neuron*

system. Mirror neurons are a class of visuomotor neurons that discharge both when an individual performs a transitive action, for example grasping something, or when it observes somebody doing the same action (Jeannerod [1994]). These neurons were at first discovered in the F5 area of the monkey brain by measurements of the activity of a number of neurons while a monkey was either performing a grasping action itself or while watching the experimenter grasping the object (e.g. Rizzolatti et al. [1996]). The results show that neurons with mirror properties discharge in both cases, when performing and when viewing an action. It has therefore been assumed that when observing an action this action is automatically retrieved in the viewer but not necessarily executed. Consequently, one can hypothesize that the monkey in fact directly perceives the action carried out.

The mirror neurons found in the F5 area show a large degree of generalisation. They behave alike no matter whether the action is performed by a human or another monkey, independently of the kind of object involved and independently of where the action takes place, whether close to the monkey or further away (Rizzolatti and Craighero [2004]).

To go even further, it was found that the mirror neurons discharge when the monkey cannot see the action but knows that the action is performed (Umiltà et al. [2001]). In the supporting experiment the monkey saw a grasping action performed by the experimenter. During the second half of the experiment the final part of this action (when the object was grasped) was hidden from the monkey. Since the monkey knew, however, that there was an object to be grasped, the mirror neurons still discharged.

The F5 area in the monkey brain contains neurons related to hand actions, mouth actions and communicative actions. It has therefore been seen as the homologue to Broca's area in the human brain containing neurons related to the same actions (Rizzolatti and Arbib [1998]). Related to communication, the mirror system could be the link between the actions of the sender and the percept of the receiver: The receiver directly perceives the actions carried out by the sender in form of motor representations. During speech it is therefore possible that listeners perceive the gestures performed by the speaker. Fadiga et al. [2002] have carried out an experiment supporting this. They have measured motor evoked potentials (MEPs) when human subjects were listening to words, pseudo words and bitonal sounds. The words and pseudo words contained the labiodental fricative /f/ or the English continuant /r/. For the production of the first sound no tongue movement is necessary whereas for the production of the second sound tongue movement is necessary. The results show that there is an increase in MEPs recorded from tongue muscles when the subjects listened to the words and pseudo words containing /r/, but there is none in all other cases. This suggests that speakers perceive a

motor representation while listening to the sounds.

Watkins et al. [2003] carried out another experiment offering further evidence. They measured motor evoked potentials in hand and lip movements while subjects listened to continuous prose and to nonverbal sounds, while viewing speech-related lip movements and while viewing eye and brow movements. Listening to speech enhanced MEPs in the lip. No differences were found for the MEPs measured for the hand in this condition.

In a study by Keysers et al. [2003] "audiovisual mirror neurons" were found in monkeys which discharge when monkeys hear sounds related to an action (e.g. paper ripping) even without seeing the action being performed. If a mechanism like this exists in humans it could be used for the direct perception of articulatory actions in communication.

Rizzolatti and Arbib [1998] have set up a theory of language development taking into account the findings on mirror neurons. According to these authors at first a closed system of manual and oral gestures developed as a consequence of a lack of inhibition of the action during perception of an action. Then an open manual gestural system, where gestures had referential function, followed. These manual gestures were then accompanied by orofacial gestures. This system later developed into a vocal system.

Both the Motor Theory and Direct Realism thus support articulatory perceptual primitives. The speaker produces the articulatory movements he wants to transmit and the listener can perceive these articulatory actions directly, possibly with the help of a mirror neuron system. The acoustic signal only functions as a transmitter.

1.2 The sound carries the message: Phoneme representations are acoustic

As indicated, there is an alternative view, namely that the linguistic information is taken directly from the acoustic signal without reference to the articulatory movements. The articulatory movements would thus only be a means to produce the acoustic signal. A number of experiments and investigations have tried to find support for the perception of acoustic properties. Some of these will be discussed now. Afterwards, speech perception theories based on acoustic perceptual primitives will be described.

One piece of evidence for acoustic representations comes from sound change. Ohala [1996] describes a case of "spontaneous nasalisation" which means nasalisation of vowels although there are no nasal consonants next to these vowels. The cause for this nasalisation is, according to Ohala, the high

airflow elements (fricatives, affricates, aspirated stops) by which the vowels were surrounded. During their production the glottis is wide open. This glottal opening spreads into the vowels, and since thereby acoustic conditions are created which mimic vowel nasalisation speakers misperceive the vowels as nasalised. For Ohala this result shows that the object of perception is not the velar gesture (which is actually not there in the beginning) but the acoustic signal (lowered amplitude and increased bandwidth of F1).

Further evidence against gestural representations comes from an experiment with six Japanese quail (Kluender et al. [1987] and Kluender [1991]) which were trained to distinguish voiced and voiceless stops according to differences in voice onset time. The birds were able to distinguish the sounds similar to humans although they have no human vocal tracts to produce them in the way humans do and therefore cannot know and perceive gestures.

Similar evidence comes from the investigation of glossectomee speech. Morrish [1990] describes the speech of a patient with 100% tongue removal. This speaker produced speech so that 58% of the words were correctly understood. The speaker managed to do this by choosing completely different articulatory configurations as compared to normal speakers. For example, in order to produce a stop which was perceived as an alveolar one he produced a bilabial closure with lip protrusion. Even if the speech of this patient acoustically differed fundamentally from the speech of normal speakers the differences in articulation are much greater. There are nearly no similarities between the articulation of this speaker and normal speech. Still, listeners were able to recognise the phonemes the speaker intended to produce. This is hard to explain when listeners are assumed to perceive articulatory actions.[2]

Further evidence for acoustic perceptual primitives comes from a perturbation experiment described in Jones and Munhall [2003]. In this experiment the upper incisors of speakers were prolonged so that due to the thus longer front cavity the centre of gravity[3] of /s/ was lower as long as speakers did not adapt. Speakers' auditory feedback was masked during several subsessions of the experiments. The results show that speakers compensate by changing the articulation so that the centre of gravity becomes higher again. However, the compensation does not start until auditory feedback becomes available. The result shows that under perturbation speakers adapt towards acoustic perceptual primitives even if they have to change their usual articulation in order to do so.

There are a number of studies which have compared variability in ar-

[2]Fowler [1990] explains the perception of this speaker's speech as perception of a mirage: A different articulatory event causally structures air in the same way in which a known articulatory event would structure it.

[3]average of frequency and intensity components of the spectrum, cf. section 4.7.3

ticulation and acoustics of a certain speech phenomenon in order to infer something about the nature of speech production tasks. The basic idea behind these studies is that in the domain where the representations can be found differences between phonemes should be maximised. Allophonic differences, however, should be minimised. One piece of work in this tradition is the one by Lindblom and colleagues (Liljencrants and Lindblom [1972] and Lindblom [1986]) who state that if representations were acoustic, phonological systems should be built in a way so that the acoustic contrast between the sounds is maximised. The degree of difference in articulation, however, should not matter. As evidence for a maximisation of contrast in the acoustic domain Lindblom and colleagues discuss languages with five vowels which usually exhibit /a, e, i, o, u/. These sounds are distributed equidistantly over the acoustic space. If representations were articulatory the set of respective unrounded back and rounded front vowels should be equally common, but in fact the second set is much rarer. Lindblom states that the reason for this asymmetry is that the acoustic difference between the vowels of the first set is greater than between the vowels of the second set.

Another way to compare articulatory and acoustic variability is to look at *motor equivalent strategies*. Motor equivalent strategies are articulatory strategies which differ but which result in the same acoustic output. For example, the American English /r/ can be produced with tongue bunching or as a retroflex (Westbury et al. [1998], but see Zhou et al. [2007] for acoustic differences). There are thus two possible articulations, but the differences in the acoustic output are minor. Since speakers really make use of both of these strategies this supports acoustic perceptual primitives since the acoustics stay constant whereas the articulation varies. Similarly, motor equivalent strategies can also be found in /u/. This sound is usually produced with a labial and a velar constriction. Speakers can widen the velar constriction and acoustically compensate for that by more lip protrusion. Perkell et al. [1993] found for EMA data that some speakers in fact use this relation and covary the two parameters.

Another study dealing with motor equivalent strategies in /u/ is the one presented in Savariaux et al. [1995] and Savariaux et al. [1999]. During these experiments the lips of subjects were hold open by a 2 cm diameter tube and the subjects were asked to produce a normally sounding /u/. After a training phase some subjects produced nearly the normal F1-(F2-f0) pattern of /u/ either by creating a velo-pharyngeal instead of a velar constriction or by moving the constriction backwards in the velar region in association with an increase of f0. By changing the articulatory strategy in order to reach a certain acoustic output some subjects thus show great changes in the articulatory domain in order to minimise changes in the acoustic domain.

The last piece of evidence which will be discussed before turning to the acoustic theories comes from Ohala. With respect to language learning it seems to be the case that learners can perceive speech sounds and distinguish them from other speech sounds without being able to produce them themselves (Ohala [1996]). This suggests that they have an acoustic representation in mind and not the gestures or some other kind of articulatory representation which should make it easy to produce the sounds. According to Ohala, imitating the sounds seems to be a trial-and-error process where speakers try to find the gestures matching the acoustic representation of the sound.

The reason for reducing the within-category variability is probably to enable the listener to catch the core characteristics of the category. That is why the first perceptual theory claiming that perceptual primitives can be found in the acoustic domain, the *Theory of Acoustic Invariance* (Stevens and Blumstein [1978], Blumstein and Stevens [1979]) was interested in finding invariants. The theory states that for every phonologically distinctive feature there is an invariant property, either it is in the acoustic signal itself or it arises during the processing of the signal. Examples for invariants have been claimed to be the formant patterns of vowels and the spectral shape of the burst of stops. In line with the proposals by Jakobson et al. [1961], Stevens and colleagues developed a model of speech perception called *Lexical Access from Features* (Stevens [1986], Stevens [1988], and Stevens et al. [1992]) claiming that words are stored in the brain as patterns of these distinctive features and that these features are associated with acoustic invariants.

For some of the features evidence for invariants in the acoustics could be found. Zue [1985], for example, found a set of "robust features" in the acoustic signal which seem to be independent of speaker, context and speaking style. These features include [fricative], [nasal] and [voicing]. For most other features, however, invariants could so far not be found, at least not across speakers, context and style. Moreover, Jessen [1999] even showed that, from a cross-linguistic perspective, there are no invariants in the voicing contrast, instead there are a number of characteristics which can be used to distinguish between voiced and voiceless sounds, and languages differ in their choice among them.

Miller [1989] extended the idea of invariants to invariant target regions. For American English vowels he found these regions in a three dimensional "auditory-perceptual space" which is calculated from the fundamental frequency and the first three formants. Within the three dimensional space Miller defined separate regions for American English vowels.

Since convincing evidence for invariants corresponding to all features has not been found so far, later theories focussing on acoustic representations

try to find the invariants elsewhere than in the acoustic signal. The *Auditory Enhancement Theory* (Diehl and Kluender [1989]) assumes that acoustic properties combine to form "intermediate perceptual properties" (IPPs) which again combine to form distinctive features, which are invariant. The phonological distinctions of the language will be enhanced by the process of combining. What children learn when learning a language is to combine the acoustic properties of the language in order to enhance features. Support for this theory has been found by trading relations in categorical perception (Diehl and Kingston [1991]), e.g. in voicing. [+voice] judgments increase for low F1, low f0 and longer voicing durations in stops. Diehl & Kingston therefore suggest that these three parameters are merged into the perception of only one characteristic. In contrast to Fowler's conclusion from these trading relations, that there is one laryngeal gesture which has several auditory effects, e.g. a certain f0 or a certain segment duration, Diehl and Kingston think that these auditory effects are independent from biomechanics. Taken together, they enhance the feature *voicing*.[4] Kingston and Diehl [1994] even claim that f0 changes for voiced vs. voiceless sounds are not a consequence of a laryngeal gesture (as proposed by Fowler) but they are independent of this gesture. In order to support this claim they discuss an example of a voiced singleton and a voiceless geminate in Tamil. For both these stops f0 at the onset of the following vowel is essentially the same and can therefore, according to Kingston & Diehl not depend on the laryngeal gesture.[5]

Kingston and Diehl [1995] again dealt with the [voice] feature and tested the two principle cues (IPPs in the terminology of the Auditory Enhancement Theory) which contribute to the feature [voice] set up by Stevens and Blumstein [1981]. These two cues are (1) the CV duration ratio and (2) the low-frequency property. The first cue or IPP is put together from two acoustic properties, namely preceding vowel length and closure duration. The second

[4]This approach is, however, very controversial, see Löfqvist et al. [1989] and Hoole [2006] for alternative approaches to the relationship between voicelessness and raise in f0. Loefqvist et al. compared CT activity in voiced and voiceless sounds and found that CT activity is higher during voiceless sounds. The higher f0 at the onset of the vowels following voiceless stops is explained with the long relaxation time of the CT so that the vocal folds are still tenser after voiceless sounds than after voiced sounds. Hoole [2006] argues that the reason for the increased CT activity is not carried out to suppress vocal fold vibration but to increase the glottal opening by supporting the abductory motion of the arytenoids. The fact that the CT traces of voiced and voiceless stops stay separated right through the duration of the following vowel, however, suggests that speakers are enhancing a contrast between voiced and voiceless stops which exists already due to the glottal mechanism.

[5]Hoole [2006], however, argues that in the geminate the abduction-adduction cycle is completed earlier with respect to the following vowel onset than in the singleton so that the influence on f0 has already diminished when the vowel following the geminate starts.

cue or IPP is put together from voicing into closure, F1 at the boundaries of the closure and f0 at the boundaries of the closure. In general, a long preceding vowel, a short closure duration, a long voicing into closure, a low F1 and a low f0 are related to the perception of a voiced stop. All these properties were varied in synthetic non-speech stimuli in order to assess whether the enhancement is really an auditory property and is not learned during speech acquisition. Listeners were asked to judge whether the "gap" (the non-speech equivalent to the stop closure) is long or short. The results show that all the five characteristics contributed to the perception of gap length. Kingston & Diehl thus suggest that speakers merge these acoustic characteristics in only one percept.

Another theory dealing explicitly with variability is the *Adaptive Variability Theory* (Lindblom [1988] and Lindblom [1990]) according to which variability is a part of speech and the degree of variability produced depends on the communicational situation. The speaker produces as much variability as the listener can tolerate. Both speaker and listener have information about the communicational situation and the listener factors the communicational situation into his "calculation" of the meaning of what is said. This means that listeners get all the essential information from the acoustic signal and the situation, the invariants, however, can only be assumed in higher-level processing stages of the signal.

1.3 Conclusion

Evidence for both acoustic and articulatory perceptual primitives has been found. Invariances in the acoustic signal have so far not been found. A mechanism which could enable listeners to directly perceive articulatory actions could be the mirror neuron system.

A conclusion which could be drawn is that maybe both exist, articulatory and acoustic properties in the speech production tasks. The experiment described in Jones and Munhall [2003] has shown that speakers are able to produce speech when the auditory feedback is masked. Under these circumstances speakers could use an articulatory representation. When auditory feedback becomes available, however, this articulatory representation seems to be overrun by the acoustic representation. In order to keep the acoustic properties of the sound speakers in this experiment change the articulation. It could thus be that articulatory representations are pure motor representations which exist for the speaker but they are not transmitted to the listener.

The experiment discussed in this study shares some basic properties with the Jones & Munhall experiment. Speakers' articulation is perturbed and

speakers adapt. The feedback is sometimes masked in order to investigate whether speakers need it in order to adapt. The present experiment differs, among other things, with respect to the kind of perturbation. When the palate shape is changed speakers should be able to adapt with the help of tactile feedback (which should have been difficult in the study by Jones and Munhall). Thus, in contrast to this earlier study speakers in our study can be expected to compensate even without auditory feedback available and it should be interesting how far this compensation goes.

Chapter 2

Motor aspects of articulation: Movement optimisation

From the discussion in the preceding chapter the conclusion can be drawn that under perturbation speakers' aim should be to produce an acoustic output which either carries the linguistic message itself or which provides information about the articulatory gestures carrying the linguistic message. As mentioned already, this study does not deal with perception but with production of speech. That is why in addition to perceptual targets it has to consider movement organisation principles because they could interfere with the primary aim of the adaptation, to transmit the *common currency* (cf. p.14) from the speaker to the listener. For example, apart from this primary aim, and possibly independent of it, speakers should, with more and more practice, try to reduce the articulatory effort involved in their productions. In contrast to the first aim this is not special to speech, but it applies to human movements in general. Emphasis will now be placed on this second aim during the adaptation, and a number of previous studies dealing with characteristics of human movements (limb and articulator movements) will be discussed. The focus is laid on two approaches, one by Nelson and one by Jordan, since these two approaches will be of importance for the discussion of movement optimisation later on.

Physical correlates of "little effort" or "optimal" movements have been seen in several measureable parameters, for example short movement duration, small movement distance and the "smoothness" of the movement. In most approaches movement optimisation is seen as being subject to limits imposed by fixed constraints, either due to the *task* or the *system* involved. To give an example for a task constraint, whereas a reaching movement with a small movement amplitude might involve little effort, it might not be suitable in order to reach a spatial target. In order to fulfill the task, the movement

amplitude therefore cannot be arbitrarily low. An example for a system constraint is that every biological system has a maximal velocity it can reach. Thus, the velocity cannot be arbitrarily high. Assuming more than one criterion, a further limitation is that sometimes the criteria are not compatible with each other. For example, a quick movement (optimal duration) is not necessarily smooth.

While the concept of *smooth movements* is intuitively clear, several attempts have been made in order to quantitatively assess the smoothness of a movement. The most common concept of a smooth movement is the *minimum jerk* trajectory (e.g. Hogan [1984], Nelson [1983], Jordan [1996]). The jerk of a movement is the third derivative of it and thus gives information about the changes in acceleration during the movement. Very broadly, a minimum jerk trajectory is characterised by a single velocity peak, a single acceleration and a single deceleration phase. Other ways to assess smoothness are the minimal peak acceleration trajectory, the minimal force, minimal energy (Nelson [1983]) and minimal torque change trajectories (Uno et al. [1989]).

Hogan [1984] presents a model generating movement trajectories which are subject to a minimum jerk criterion. The results of simulations carried out with the model were compared to data gained for monkey forearm movements and turned out to be similar. With respect to the reasons for why the monkeys should produce minimum jerk trajectories Hogan states that they try to prevent moving to the limits of neuromuscular performance and thus "minimize the "wear and tear" on the neuromuscular system" (p.2751).

Whereas Hogan thus describes only one movement optimisation criterion, the approach by Nelson [1983] distinguishes several parameters and furthermore imposes constraints. A constraint, as discussed above, can be the maximally possible acceleration which a system is able to produce. Other constraints can be the movement amplitude which is needed in order to fulfill the task or the time which is given to fulfill the task. According to Nelson, optimisation can be carried out via a reduction of *force, impulse, energy* and/or *jerk*, but also - if these are not constraints - by *time* and *amplitude* reduction.

From these parameters Nelson defines *costs*, which can be regarded as a kind of measurement of articulatory effort. Very generally, one can say that the force cost is high for high peak accelerations, the impulse cost is high for high velocities, the energy cost and the jerk cost are high for high changes in acceleration as they can be found in movements which are not smooth.

In Nelson's approach the different objectives of the movement criteria are sometimes not compatible with each other, for example, a minimisation in time leads to an extreme rise in force, impulse, energy and jerk. Minimum

impulse movements, on the other hand, always lead to a rise in the jerk and energy costs. The influences the parameters have on each other are, however, not linear. This is demonstrated in the article by a discussion of the relation between energy and time. Minimal time for a movement with a fixed movement amplitude leads to a very high energy cost. Increasing time leads to a sharp drop in energy up to a point where a further increase in time no longer leads to a considerable reduction of the energy cost. For experimental data Nelson found that the kinematic characteristics of jaw movements during speaking can be found at this point where the energy cost is rather low for a reasonable time cost. For violin bowing on the other hand, Nelson found a minimum impulse pattern, probably in order to reduce differences in velocity which would work against fulfilling the task. With respect to the reasons for an organisation of the movement with the aim to reduce movement costs, Nelson, in a similar way as Hogan, states that humans (or animals) try to avoid operating at the limits of one or the other objective.

A further observation by Nelson is that minimum jerk trajectories are very similar to the trajectories which are produced by an undamped linear spring model. This could lead to the assumption that the production of minimum jerk trajectories is inherent to the muscle system. However, the fact that different findings were reported for violin bowing as compared to jaw movements supports the view that cost minimisation as such is not inherent to the muscle system.

In contrast to Hogan, who has concentrated on arm movements, and Nelson, who deals with arm and jaw movements, Jordan [1996] deals with articulator movements in general. An important contribution from this approach is that the limits within which movement optimisation is carried out are no longer seen in purely spatial terms. Jordan defines a *task constraint* which ensures that the task (to reach an acoustic target) is fulfilled. Three further *generic constraints* can be used for optimisation. The first of them is the *smoothness constraint*, which aims at producing a minimal jerk trajectory. The second constraint is the *distinctiveness constraint*. Its aim is to produce sounds which are maximally distinctive from each other. The third constraint is the *rest-configuration constraint*, which can be seen as an attempt to minimise articulatory effort by moving as little as possible from a neutral articulatory rest position. From these four constraints a cost functional is calculated within which different weights can be given to each generic constraint. Optimisation is seen as a learning process during which this cost functional is minimised.

Jordan, in a similar way as Nelson, argues that the constraints work against each other. For example, the distinctiveness constraint might require

that an articulator moves very far away from the rest position. The rest-position constraint on the other hand, requires the articulator to stay close to the neutral position.

For his speech production model Jordan describes a learning process with two aims. The first aim is to produce the required acoustic output and the second is to produce an optimal articulatory trajectory while minimising the cost functional. Jordan's speech production model is an *internal model* approach. Thus, a main assumption in this approach is that speakers dispose of a model of their own speech production system. This internal model consists of two components, a forward model and an inverse model. The *forward model* contains a mapping from motor commands to acoustic outputs. It can thus predict the acoustic result of a given motor command before the motor command is executed. The forward model furthermore predicts the kinematic properties of the movement (for example the smoothness). The second component of the internal model is the *inverse model*. This model uses the same mapping between motor commands and acoustic output, but in contrast to the forward model, which predicts the acoustic output, the inverse model goes the other way and tries to find a set of motor commands which produce a desired output. Since the mapping between motor commands and acoustic output is many to one (many sets of motor commands can produce the same acoustic output) the inverse procedure, which proposes a set of motor commands for a certain acoustic output has to involve a selection process. This selection process is carried out according to the optimisation criteria just mentioned (while using the kinematic properties provided by the forward model). A particularity of this process is that the model tries to find an optimal movement already when it has not yet succeeded in finding the correct output.

The studies discussed in this chapter suggest that the jerk of a movement is a rather good indicator of articulatory effort. The speakers of the present study might, with more and more practice, produce movements which are smoother and smoother over the adaptation time. Thus, for the data presented here it should be interesting to investigate under which circumstances movement optimisation takes place and in how far it interacts with the main aim of adaptation, namely to reach the phoneme targets.

Coming back to the distinction between *perceptual primitives* and *speech production tasks*, movement optimisation belongs only to the second concept, but not to the first one. When speakers adapt towards a perturbation, what one can see is a development with the aim to fulfill a speech production task. Consequently, if one wants to conclude something about perceptual primitives one has to make sure that it is not something from the domain of movement optimisation. For example, if speakers show very little variability

in articulation towards the end of the experiment this might not necessarily mean that they try to transmit articulatory information to the listener. It can also mean that they have selected the most optimal articulatory strategy leading to a certain acoustic output.

Chapter 3

Speech perturbation and compensation

The discussion in the preceding two chapters turned around the primary and the secondary aim of adaptation towards a perturbation, respectively. The primary aim should be to produce speech which transmits the linguistic message. The secondary aim should be to produce speech with a minimum of effort. Before turning to the experiment, the introductory part of this study will be concluded by an overview of previous perturbation experiments. The aims of this overview are (1) to set up and justify the methodology for the present experiment and (2) to develop expectations with regard to the behaviour of the subjects in the present study. In these earlier studies, depending on the type of the perturbation, speakers showed the three different kinds of compensatory strategies discussed below. For introductory purposes at first a brief definition of each strategy is given together with an example. Further below each strategy will be discussed in detail and experimental examples will be described.

- For a great number of perturbations, speakers apply a *stabilisation strategy*: They try to reduce the influence of the perturbation without changing the underlying movement pattern. For example, when producing speech in different bodily postures, speakers have to control their jaw muscles differently. While standing, the jaw can be opened without much force of the sternohyoid muscle due to gravitation, whereas in a supine position more active force of this muscle might be used.

- A second type of compensatory behaviour is the *reparametrisation strategy*. As in the stabilisation strategy, speakers here rely on a learned pattern but assign different values to parameters which are part of the structure. For example, speakers manage to speak while keeping the

jaw in a fixed position, e.g. with a cigarette in the mouth. Here different degrees of involvement in gesture formation are assigned to jaw and tongue.

- The third type of compensatory behaviour is the set up of a new articulatory strategy. This type of compensatory behaviour will be called *reorganisation strategy*. It is used when speakers have to adapt to a structural change of the vocal tract shape. For example, when people speak with food in the mouth the vocal tract shape is changed and speakers have to adapt their articulatory behaviour. Dental devices are another example for a perturbation resulting in a reorganisation. Whereas stabilisation and reparametrisation occur immediately, the third type of compensation normally requires some practice.

Examples for experimental investigations for each compensation type will now be discussed.

3.1 Compensation via a stabilisation strategy

Many perturbations can be compensated for by simply stabilising the usual movement pattern, for example by an increased co-contraction of the muscles. As mentioned above, this kind of strategy is used when speaking in different bodily postures. Shiller et al. [1999] investigated speech in upright, supine and prone position. Each posture involves the exertion of different gravitational loads on the jaw. The resulting position of the jaw in the different conditions was investigated. The main finding of the study was that jaw positions differed for the three postures, except for the horizontal position of the jaw (anterior-posterior) which was the same in the upright and supine orientation. The compensation is thus not complete in all cases. Shiller et al. compare their results to the ones of arm movement studies. In this kind of study subjects are asked to keep a certain limb posture. A force is then applied to the limb so that the posture changes. While compensating, subjects reestablish the initial posture. The differences found between jaw and arm movements are assumed to be due to the different tasks. In contrast to arm movements, in speech the task is not to achieve a certain position but to produce intelligible utterances. For the study by Shiller et al. this could mean that an equal position of the jaw is not required as long as other articulators are used to produce the required acoustic output.[1] A comparison

[1] If this had been found this would have been a reparametrisation according to the classification carried out here.

of F1 and F2 values of the productions in the different positions, however, suggested that speakers did not compensate with other articulators either. The values differed significantly for the three postures. Thus, compensation for differences in gravitational load exists, but it is not complete. The compensation strategy used seems to be a pure stabilisation strategy without reparametrisation.

Tiede et al. [2000] compared articulatory trajectories of the lips and the tongue in upright and supine condition via EMA and found that they differed. There were smaller differences between trajectories for sounds with acoustically sensitive targets, e.g. /i/, than for other sounds. The acoustic differences between the productions in upright and supine posture were not significant. In a perceptual study listeners were able to distinguish between the two conditions, but only slightly above chance level. Even if the results of the acoustic measurements presented in Tiede et al. [2000] and Shiller et al. [1999] are not the same the differences in formant values seem in all cases to be rather small. They can be expected not to cross phoneme boundaries so that a phoneme identification is in all cases possible.

Shiller et al. [2001] investigated compensation for the effects of head acceleration on jaw movements while subjects walked or ran on a treadmill. Jaw movements were observed via Optotrak[2]. When subjects were not speaking the jaw movement was dependent on the head acceleration caused by locomotion. Upward head acceleration produced a downward load on the jaw, and as a consequence the jaw position was lowered. Downward head acceleration produced an upward load on the jaw. The jaw position was consequently raised. Contrary to these jaw movements in the non-speaking condition, the jaw motion was stabilised to account for the effects of head acceleration when subjects were speaking. Shiller et al. suggest that speakers use higher levels of muscle co-contraction (greater stiffness) in order to reduce the effect of head acceleration on the jaw. An alternative explanation raised is that subjects explicitly adjust the time-varying control of the jaw.[3]

A stabilisation strategy can also be assumed for the speakers in the experiment carried out by Tremblay et al. [2003] who investigated compensation for a perturbation of the jaw movement by a robotic device. During jaw movement a velocity-dependent force was applied to the jaw which pulled it towards the front. In vocalised speech and non-vocalised silent speech, initially, the jaw trajectory changed, but after some time speakers adapted and produced nearly the original trajectory. In non-speech movements, however, speakers did not adapt.

[2]motion measurement system that tracks infrared position sensors attached to a subject

[3]This would no longer be a stabilisation but a reparametrisation strategy.

Stabilisation of a learned movement pattern is thus a frequently used compensation strategy which can be applied immediately and thus does not require training.

3.2 Compensation via reparametrisation of a learned movement strategy

This second kind of compensation strategy relies, similarly to the first one, on a learned strategy. In contrast to the first one, however, certain parameters of the learned strategy are readjusted. A further contrast to the stabilisation strategy is that the aim of the reparametrisation strategy is not to produce the *same* movement as in the unperturbed condition but a *functionally equivalent* movement. Since the fundamental pattern already exists the adaptation is immediate or at least does not take very long.

The most common compensation experiment with a reparametrisation is speech with a bite-block. In this kind of experiment speakers bite on a small block which is placed between an upper and the corresponding lower molar. The jaw is thus fixed and all changes in vocal tract configuration have to be carried out by other articulators. An early example for a bite block study is the experiment by Lindblom et al. [1979] who investigated Swedish vowels produced with two different bite blocks by six speakers. They found that speakers compensated immediately so that the majority of productions had formant patterns which fell within the ranges of variation observed for normally spoken vowels. The productions do not improve any further over the adaptation time. Speakers in this experiment compensate for the unusual jaw positions by different tongue movements. Thus, the articulatory strategy is not changed as such and still the same articulators are involved but different weights are given to parameters: More weight is given to the tongue, and less weight is given to the jaw.

Kelso and Tuller [1983] investigated the spectral characteristics of the vowels /a, i, u/ surrounded by /p/ when produced in a number of conditions (bite block, no bite block, with auditory feedback masking and without). They found that compensation towards the bite block was immediate and complete even on the first trial. A number of other authors made similar observations (e.g. Flege et al. [1988]).

McFarland and Baum [1995] also compared speech in bite block and normal condition. They found small but significant differences in formant values in the vowels /i, a, u/ and lower values for the centre of gravity in /p, t, k, s, ʃ/ in the bite block condition than in the normal condition. The formant

values improved a little after a 15 min practice period, the spectral characteristics of the consonants, however, did not improve. McFarland & Baum conclude that the compensation is, in contrast to what was stated in earlier bite block studies, not complete and not immediate, and that for fricatives a longer adaptation period is needed. A perception experiment presented in Baum et al. [1996] showed that the quality of both vowels and consonants was reduced in the bite block condition as compared to the normal condition.

Folkins and Zimmermann [1982] carried out a perturbation experiment where the lower lip was pushed down unexpectedly by an electrical stimulation. The three speakers investigated in the study showed active compensatory behaviour in bilabial stop production by moving the upper lip further down and by moving up the jaw. This compensatory behaviour did not result in additional gestures but in an enlargement of the existing gestures. For example, when perturbation onsets differed with respect to an articulatory gesture, compensation did not occur relative to the perturbation onset but relative to the articulatory gesture. The results show that the underlying strategy stays the same (otherwise there should be differences with respect to different perturbation onsets) but it is reparametrised: More weight is given to the upper lip and the jaw.

Gracco and Abbs [1985] perturbed the movement of the lower lip during the production of a bilabial stop. In their experiment the lip was lowered with a paddle. Speakers compensated via an increase in movement amplitude, velocity and movement time of upper and lower lip. It was found that lower and upper lip react differently in dependence on the time lag between perturbation onset and movement onset. This time lag was measured with respect to the onset of the EMG signal of a lower lip muscle, the orbicularis oris inferior (OOI). If this time lag was comparably large (the perturbation started a rather long time before the beginning of the OOI signal) the greatest part of the compensation was carried out by the lower lip which moved either for a longer time or quicker. When the time lag was short or negative (the perturbation started after the onset of the OOI signal) the upper lip moved more, and the movement time or the velocity of the upper lip increased. Gracco & Abbs interpret the two different compensatory strategies as two different control strategies: The lower lip response is an autogenic action whereas the upper lip response is a non-autogenic one. In both cases, however, compensation is carried out against the background of the already learned strategy.

Reparametrisation strategies show that speech tasks cannot be defined in absolute spatial terms. The aim of speech production is not to produce a certain articulator position. Rather, a speech task could be defined for example as a "tract variable" (in task-dynamics terminology, cf. Saltzman

and Munhall [1989]), for example as a constriction or a closure, or in terms of a certain acoustic output. A certain degree of variability, both in acoustic and articulatory terms, is tolerated by the speakers.

3.3 Reorganisation of the movement

Reorganisation of the learned strategy is most likely to occur when speakers are confronted with structural changes of the vocal tract shape, for example the palate shape or the teeth. Confronted with this new environment speakers often cannot rely on a learned strategy which is just reparametrised but they have to develop a new strategy. Compensation therefore often takes rather long.

Earlier examples of structural perturbation experiments are the studies by Hamlet and Stone [1976] and Hamlet and Stone [1978]. Hamlet and Stone [1976] investigated compensation for a palatal prosthesis changing the palatal contour via ultrasound and acoustics in vowel production. Each subject was provided with one of three different prostheses: one which thickened the alveolar ridge by 4 mm, one which thickened the palate bilaterally (3 mm) and one which thickened the palate unilaterally (4 mm). They recorded 6 speakers, first without prosthesis, then immediately after insertion of the prosthesis, after a 15 min conversation, after one week of adaptation during which the subjects had worn the prosthesis during day time, immediately after taking out the prosthesis and 15 min after having taken out the prosthesis. Acoustic analyses and analyses of jaw and larynx movements were carried out. The compensation methods were speaker related and did not depend on the form of the prosthesis. Compensation in vowel production was not always complete. However, the acoustic differences which remained after one week of adaptation were rather small and the subjects did not notice them. The authors suggest that consonants, which were not investigated in this study, are more influenced by the prosthesis and harder to compensate for. After removal of the prostheses the vowel formants did not immediately reach the original values. This after effect shows that a new articulatory strategy must be involved.

Hamlet and Stone [1978] carried out an electropalatographic study and investigated alveolar consonant productions of ten speakers. For each speaker three kinds of artificial palates were made: (1) a normal thin EPG palate, (2) a palatal prosthesis with a built up in the palatal region without EPG electrodes in order to wear it in between the recording sessions, and (3) an EPG palate similar in shape to the palatal prosthesis. Jaw movements were recorded via a strain gauge system. Speakers were recorded when they

first attempted to speak with one of the EPG palates (thick and thin EPG palate), after two weeks of adaptation to the thick palate, and after an additional month during which no prosthesis was worn. The results show that there is tongue overshoot shortly after perturbation onset which results in more contacts in /s, z, t, d, n/ and /ʃ/. After two weeks of adaptation the number of contacts had decreased. For the alveolar fricatives the groove size is reduced when the thick prosthesis is worn for the first time. In some cases the tongue overshoot in the alveolar fricatives was severe so that a stop was produced. One compensation method was a change in place of articulation (tongue retraction or advancement). Some of these changes were still present after the end of the adaptation time. Speakers reported that the adaptation of the alveolar fricatives was not complete in the two weeks. For the stops and the nasal they experienced their compensation as complete even if more linguo-palatal contact could be found than in the normal condition. The authors reason that "the area of contact may not be a critical feature of the production for /t, d/ and /n/" (p.238). Jaw position was in most cases lower when the thick prosthesis was worn. There was a trend to have less jaw movement with prosthesis than without. The authors suggest a shift in "co-ordinate system for jaw activity" (p.241). Coarticulatory effects of jaw movement remain preserved over the adaptation time. The articulatory variability stays the same over all sessions. Again, the after effect, but also the fact that compensation took so long suggest the set up of new articulatory strategies. The fact that there was a difference in immediate compensation and long-term compensation furthermore suggests that at first speakers try out a reparametrisation strategy, only afterwards they carry out a reorganisation of the articulatory movement.

McFarland et al. [1996] present results from a short term perturbation experiment with artificial palates of two different thicknesses (3 mm and 6 mm) behind the alveolar ridge. Vowels (/i, a, u/), stops (/p, t, k/) and fricatives (/s, ʃ/) were produced without the artificial palate, immediately after insertion of the palate and after a 15 min conversation. Whereas the changes in vowel acoustics were minor, fricatives were severely affected. /t/ was also affected by the prosthesis. A perception experiment confirmed the acoustic measurements. Quality ratings of the fricatives and /t/ were rather low. A slight improvement could be found after the 15 min conversation. The authors interpret this last finding as evidence for the use of sensory feedback in the adaptation of articulatory gestures. The results are compared to a previous bite block study and it is reasoned that the compensation methods in the two experiments differ. A fixed jaw seems to perturb vowels as well as consonants whereas the presence of an artificial palate is problematic predominantly in consonant production.

Baum and McFarland [1997] carried out another palate perturbation experiment with an artificial palate of 6 mm maximal thickness at the alveolar ridge, but this time they only investigated /s/ and they provided intensive training in the breaks between the recording sessions. Speakers' productions were recorded acoustically, at first without the artificial palate, then with the artificial palate, after 15, 30, 45 min (with artificial palate in place) and after 60 min (with and without the artificial palate). The seven speakers of the study were able to compensate: Whereas in the first perturbed session the spectral mean was lower as compared to the unperturbed session, it afterwards increased without however reaching the initial value. A perceptual experiment showed that the quality ratings of the last perturbed productions were higher than the ones of the first perturbed session. The quality of the post-perturbed productions (without artificial palate), however, decreased as compared to the initial unperturbed productions. This shows that the perturbation had an effect on the unperturbed productions as well. A further aspect which is demonstrated by this experiment is that a certain degree of adaptation can be achieved rather soon if efficient training is provided.

In the study presented in Baum and McFarland [2000] the same kind of prosthesis was used, but this study concentrated on the effect of vowel context on the degree of adaptation in /s/, on the after effect on unperturbed productions and on the ability to recall the learned strategies after an hour of normal speech. Four speakers were recorded without the artificial palate, immediately after perturbation onset, after an hour of intensive practice with the prosthesis in place, after having taken out the prosthesis, after another hour of normal speech without the palate and finally with the palate in place. The authors observe a striking amount of individual differences in compensatory abilities. Intensive practice of /s/ had no effect on the production of a similar sound (/ʃ/). Speakers furthermore had difficulties to recall the learned patterns after an hour of normal speech. Negative after effects were only found for one speaker. The production of /s/ followed by /i/ was more difficult under perturbation than when the fricative was followed by /a/ or /u/.

Aasland et al. [2006] again investigated /s/, but this time via electropalatography. Two EPG palates were made for each speaker, one normal one ("thin palate") and one with a built up of 6 mm in the alveolar region ("thick palate"). Speakers were recorded immediately after insertion of each EPG palate, and after 15, 30, 45 and 60 min. In between the sessions speakers read /s/-laden passages with the thick palate in place in order to practice the production of /s/ with the prosthesis. Additionally, speakers were recorded acoustically without an artificial palate before and after the EPG experiment. The results show that the centroid frequency initially decreases for both

palates but then increases over the adaptation time. The compensation is better for the thick palate than for the thin one. Friction duration increases over the adaptation time. This increased duration is retained in the final acoustic recording. Overall there is more linguo-palatal contact for the thin than for the thick prosthesis. Quality ratings from a perception experiment increase over the adaptation time.

Honda and Kaburagi [2000] investigated speech compensation to an inflatable palatal prosthesis. The base of this kind of prosthesis can be compared to the thin prostheses used in the static perturbation experiments discussed so far. The built up at the alveolar ridge used for thick prostheses, however, was replaced by a balloon which can be inflated and deflated by the experimenter. Two speakers were provided with such an inflatable palate. Their tongue movements were recorded via EMA. Auditory feedback was temporally masked. Compensation in the sounds /t, ç, s, ʃ/ was tested. The quality of the productions was rated in a perception experiment. The palate was inflated randomly at any one of 10 repetitions of an utterance and left inflated for six repetitions afterwards. The authors distinguish between compensation to *unexpected* perturbations and *immediate* compensation. An unexpected perturbation occurs in the first one of the six perturbed trials within a row of 10 trials. From the third repetition onwards one has immediate compensation. Thus, in immediate compensation speakers have already uttered two trials and can use the knowledge gained while doing this. In unexpected perturbations speakers do not have this knowledge available. There were two conditions: With auditory feedback masked and with auditory feedback available. In the unexpected condition there was often tongue overshoot. Afterwards speakers responded by lowering the tongue blade. No compensation via the lips or the jaw could be found. Identification scores in the subsequent perception experiment were in general high except for the fricatives in the unexpected perturbation condition which were frequently classified as stops. In /s/ and /ʃ/ the identification scores were higher when auditory feedback was available than when it was not.

Honda et al. [2002] describe another experiment with the inflatable palate. Two speakers were asked to produce /ʃa/ and /tʃa/ eight times in succession in a carrier phrase. This task was repeated several times with auditory feedback masking and without. At some repetitions the palate was inflated. Later, speakers were allowed to practice speaking with the inflated palate. Then they were recorded while speaking with the inflated palate. Afterwards the palate was suddenly deflated and immediate compensation in the deflated condition was investigated. A perception experiment tested whether the productions could be correctly identified as /ʃa/ and /tʃa/. The first syllable right after inflation was normally misperceived. From the second

syllable onwards, however, the productions were usually correctly identified in the session when auditory feedback was available. In contrast to the earlier study with the inflatable palate (Honda and Kaburagi [2000]), the productions without auditory feedback available were more often misidentified in the inflation and the steady-state inflated condition because there was tongue overshoot. Deflation was less problematic. The authors assume that biomechanical saturation effects are exploited in this condition. In general, articulatory effort as well as phonation effort increased when auditory feedback was masked.

The analysis of the positional data showed that during inflation the front two EMA coils were lowered, and the back two coils moved backwards. With respect to the temporal course of the adaptation there were no differences between the masked condition and the normal condition. The reaction times were between 70 and 160 ms for the vertical tongue movement and between 280-500 ms for the horizontal tongue movements.[4] The fact that the reaction times are often very short and that a great degree of compensation is carried out during the session with feedback masking suggest that tactile feedback plays a dominant role. Auditory feedback, however, is used as well, as can be seen from the positional differences and the identification scores.

Heydecke et al. [2004] investigated different kinds of structural perturbations by comparing the influence of three kinds of dental devices on speech production. The first device was a fixed implant prosthesis which covered the palate partly by a connective part between two molars. The second and third prostheses were removable. One of them covered the palate, the other one did not. Edentulous speakers were provided with several of these prostheses one after the other. The adaptation time for each prosthesis was two months. The acoustically recorded productions were rated by two listeners. Ratings were higher for the removable prostheses than for the fixed prosthesis. The presence or absence of palatal coverage seemed to have no influence on the intelligibility. This is in contrast to what has been found in studies using for example EPG palates, and the authors explain it by the fact that the subjects had worn prostheses with palatal coverage before the experiment for several years and developed compensation strategies. The most severe perturbation was induced by the fixed prosthesis. Productions with this prosthesis were in general rated lower in quality than productions with a removable prosthesis. According to the authors a reason for this could be the space left between the alveolar ridge and the fixed prosthesis through which air passes during speech production. Vowels were in general less affected than consonants.

[4] although one could doubt whether it is real compensation or maybe rather a mechanical effect

Another perturbation experiment but where a different part of the vocal tract was modified is the one by Jones and Munhall [2003] which has already been mentioned in chapter 1. In this experiment subjects' upper incisors were prolonged and speakers' /s/-productions were investigated. Speakers were recorded alternately with auditory feedback masking and without. Initially, the centre of gravity of the /s/-productions was lowered as a consequence of the prolonged front cavity due to the longer incisors. Speakers started compensation only when auditory feedback was available. This experiment shows the importance of auditory feedback when tactile or proprioceptive feedback does not provide any information about the kind of perturbation. Whereas in the palate perturbation experiments speakers could acquire information about the perturbation before they started to speak so that they could adapt even in the first utterance, speakers in the study by Jones & Munhall could not. They had to wait for auditory feedback.

3.4 Conclusion

This section has discussed a number of previously carried out perturbation experiments. Three types of compensation can be distinguished:

- stabilisation
- reparametrisation
- reorganisation

The first compensation strategy does not involve a change of the articulatory strategy nor does it, if it is successful, involve a change of the articulatory movements. An example is the use of more co-contraction of antagonist muscles in order to compensate for different bodily postures.

The second compensation strategy involves a reparametrisation within the framework of a learned articulatory strategy. The movement thereby changes in absolute terms but the reparametrisation leads to a functionally equivalent output. An example are the bite block experiments where the tongue compensates for the perturbed jaw movement.

The third compensation strategy involves a reorganisation of the articulation and therefore the setup of a completely new strategy. In contrast to the first two types of compensation reorganisation is never immediate and often leads to an after effect due to problems of the retrieval of the old strategy.

For the experiment discussed in this study one can therefore conclude that

- The perturbation should lead to a reorganisation of the articulatory movements.

- From the experiences made in previous experiments one can assume that speakers will, in order to compensate immediately, even when auditory feedback is masked, at first try out a reparametrisation. The degree of compensation carried out by reparametrisation can be expected to be sufficient in order to correctly identify the sound.

- Speakers will try out a number of strategies over the two weeks, they will not stay with a certain early acquired strategy even if the target (acoustic or articulatory) is reached.

Same as for the studies discussed in chapter 1, some of the findings speak for articulatory representations of speech production tasks, however, it is possible that these articulatory representations have purely motor functions and do not belong to the speech percept. They thus remain with the speaker and are not transmitted to the listener. For example, the fact the speakers can adapt with auditory feedback masked suggests that they are using an articulatory representation of the sounds which could possibly be based on a certain linguo-palatal contact pattern.

Other findings suggest that speakers use an acoustic representation. In the study presented in Hamlet and Stone [1978], for example, it was found that for the adaptation towards an artificial palate in the sounds /t, d/ and /n/ the contact pattern is not important. This could be because the contact pattern hardly influences the acoustic output in these sounds. Furthermore, there are findings (Honda and Kaburagi [2000]) that for the adaptation of the alveolar and the postalveolar fricative auditory feedback is absolutely necessary.

A number of aspects are not clear, for example, why speakers retract the tongue when they are confronted with a palatal prosthesis rather than just to lower the jaw. This changes the acoustic output and the articulatory configuration as well. For palatal prostheses which move the alveolar ridge posteriorily one could say that speakers go on producing e.g. alveolar sounds and therefore retract the tongue. However, this result could also be found for other kinds of prostheses.

For the experiment which is going to be discussed here, one can thus expect to find both, adaptation towards articulatory or acoustic targets. A question will be whether one of the two occurs earlier or one predominates over the other.

Chapter 4

Methods

The experiment carried out was a long-term perturbation experiment where speakers wore an artificial palate for two weeks and tried to adapt their articulation so that their productions appeared normal to them. Speakers' articulator movements and the resulting acoustic signal were recorded via electromagnetic articulography (EMA). Very generally, there were the following recording sessions[1] which will be explained in more detail in the due course of this chapter: On the first day speakers were recorded via EMA at first without the prosthesis, then with the prosthesis and auditory feedback masked, afterwards with the prosthesis and auditory feedback available. The next recording took place after half a week adaptation and was a pure acoustic recording with artificial palate. After one week adaptation time there was another EMA recording with artificial palate. A second acoustic recording took place after one and a half weeks of adaptation. Finally, after two weeks, speakers were recorded with the prosthesis in place and immediately after taking out the prosthesis. German lingual obstruents, tense vowels and one lax vowel were recorded.

The present chapter describes the experimental method (2D and 3D articulography and acoustics, section 4.1), the morphological changes carried out (section 4.2), the speakers (section 4.3), the temporal arrangement of the recordings (section 4.4), the auditory feedback masking (section 4.5), the speech material (section 4.6) and some of the basic analysis methods used (sections 4.7 and 4.8). The main part of the analysis methods used, however, will be described in the chapters thereafter, where also the results are discussed.

[1] A summary of all sessions can be found in the list on p.48f.

4.1 Articulatory and acoustic recordings

Speakers were recorded via electromagnetic articulography (Hoole [1996a], Hoole et al. [2003]).

The data were recorded in the framework of a Berlin-Munich co-operation project. The data from Berlin were recorded in the phonetics laboratory of the Zentrum für Allgemeine Sprachwissenschaft, the data from Munich were recorded at the Institut für Phonetik und Sprachverarbeitung of the Ludwig-Maximilians-Universität Munich. For the recordings in Berlin two-dimensional EMA was used, the speakers in Munich were recorded with three-dimensional EMA.

For the 2D recordings carried out in Berlin the Carstens AG 100 (10 channels) in the phonetics laboratory of the Zentrum für Allgemeine Sprachwissenschaft was used together with the supplementary correction program and preprocessing software described in Hoole [1996b] and Hoole [1996a].

The subject was placed in a sound proof room and could be viewed by the experimenters through a window. The stimuli were presented on a screen which was placed in front of the window so that the subjects could read them. Since in previous experiments (cf. chapter 3) it was found that subjects spoke with a higher sound amplitude and increased f0 when the auditory feedback was masked, another screen with a sound level measurer was placed next to the stimulus screen so that the subjects could control their loudness.

Eight sensor coils were attached to the subject (figure 4.1). Three were glued to the tongue, one around a centimeter behind the tongue tip, one at the part opposite the border between hard and soft palate and the third one in the middle between these two. Another sensor was placed below the lower incisors in order to track jaw movements. In order to record lip movements, two further sensors were glued to the upper and the lower lip. Two sensors at the upper incisors and the bridge of the nose served as reference sensors to compensate for head movements. The remaining ninth and tenth sensors were used to record the occlusal plane after the recording of the speech material.

The experiment included recordings on several days. Across recordings the position of the tongue sensors differed slightly. This was because whereas it is easy to find approximately the same place for the sensor at for example the lower incisors (midsagittaly below the incisors), it is hard to find landmarks on the tongue which could serve as points of orientation for finding the same position in a further recording. Photos of the tongue with the sensors on it were taken and the distance between the sensors was measured, in order to find the positions, however, this method turned out to be not completely satisfactory.

A parallel acoustic recording was carried out. The start of the recording

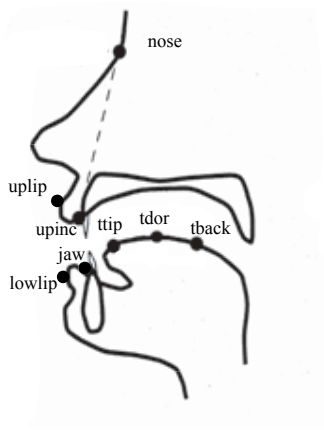


Figure 4.1: Positioning of the sensors in the 2D arrangement. Three sensors were placed on the tongue (*ttip, tdor, tback*), one at the lower incisors in order to track jaw movements (*jaw*), one at each lip (*uplip* and *lowlip*), and two reference sensors (one at the upper incisors, *upinc*, one at the bridge of the nose, *nose*.)

time for each sweep was signalled by a beep. Since the subject was not able to hear this signal during the recording with auditory feedback masking, an LED was glued to the screen presenting the stimulus. This LED gave an optical signal for the duration of the recording time so that the subjects knew when to speak during the recording with feedback masking. The sweep length was set to 2 s.

After the recording of the speech material (cf. section 4.6 for the corpus), one sensor was removed from the tongue and the contour of the palate was recorded by moving this sensor along the palate. The occlusal plane was recorded while the subject was biting on a T-piece to which two sensors had been attached midsagittally.

The acoustic signal was recorded with a Sennheiser Mkh 20 P48 microphone on the second track of a digital audio tape (44 kHz sampling rate). A synchronisation impulse generated by the PC running the EMA system was recorded onto the first track of the tape.

After the recording a number of preprocessing steps were carried out which included correction algorithms for head movement, filtering of the data, rotation and translation of the position data, synchronisation with the acoustic data and file format changes (cf. Hoole [1996b]).

The acoustic data from the DAT were fed into a computer, converted into a wav-file, downsampled to 24 kHz and cut into sweeps containing only one sentence on the basis of the synchronisation impulse.

Velocities and accelerations are calculated as described in Hoole [1996b].

Velocity and acceleration in the x dimension (VX and AX, respectively) were calculated as first and second derivative of the x movement data. The same calculation was carried out for the y dimension. Tangential velocity (VT) was calculated as:

$$VT = \sqrt{VX^2 + VY^2} \tag{4.1}$$

and tangential acceleration (AT) correspondingly:

$$AT = \sqrt{AX^2 + AY^2} \tag{4.2}$$

Velocities were low pass filtered with a cutoff frequency of 25 Hz.

For the Munich data the 3D system as described in Hoole et al. [2003], Zierdt et al. (1999, 2000) was used. The apparatus was the Carstens AG 500.

Whereas the 2D system works with three transmitters in order to extract and correct two dimensions, the 3D system works with six transmitter coils with different orientations and enables the extraction of three spatial dimensions and two orientations (azimuth and elevation).

The sensor placement was the same as for the 2D arrangement, except that there were three additional sensors, one at a mouth corner and two reference sensors laterally above the upper molars. The stimuli were presented over a screen which was placed in the recording room. Start and end of the recording time were presented visually by different colours of the stimulus on the screen.

4.2 Artificial palates

For each of the seven participants at first a dental impression of the upper jaw was made by a dentist. Then a dental technician made a model of the upper jaw. The target palatal contour was afterwards indicated by an Optosil[2] insertion into the model. Then the model together with this insertion was returned to the technician who built a palatal prosthesis according to the indication. The material of the prosthesis was the same as is used for braces, the part covering the palate consisted of acryl and the arms holding the prosthesis at the teeth were made of metal.

There were two types of artificial palates, the first one moved the alveolar ridge posteriorily ("alveolar palate"), the second one made the palate flatter by filling out the palatal arc ("central palate"). The thickness of the palates differed from speaker to speaker and depended on the anatomical conditions of the speaker. However, they all had a maximal thickness of around 1 cm.

[2]material to form dental impressions

Table 4.1: Speakers participating in the study. Column 1: initials, column 2: sex, column 3: age at time of recording, column 4: type of prosthesis (alveolar or central), column 5: laboratory (Berlin or Munich and consequently 2D or 3D EMA)

speaker	sex	age	prosth. type	lab
TP	m	39	alveolar	B
KD	f	28	alveolar	B
OP	f	28	central	B
SK	f	41	central	B
BP	m	56	central	B
AP	f	ca. 25	alveolar	M
DS	m	ca. 25	alveolar	M

4.3 Subjects

Seven speakers took part in the study. Table 4.1 gives information about the speakers. All the speakers except one (BP with a velopharyngeal dysfunction and a slight left-sided hearing impairment) were without pathological findings and all of them had a rather domeshaped palate so that there was enough space for the insertion of further material. The speakers' sensitivity towards the experimental setup was tested in advance in order to exclude participants with nausea or extreme sensitivity towards foreign substances in the mouth. Furthermore, in advance the speakers were informed in detail about what the experiment would be like. They were given the chance to look at the equipment and artificial palates of previous participants. They were not given any hints with respect to possible adaptation strategies. Participants were instructed to wear the palate all day except during eating and sleeping. One subject (BP) decided for himself to wear the palate during the night as well. The speakers from Berlin were given a sheet with speaking exercises (from Fiokowski [2004]) and they were asked to read through these exercises aloud once a day.

4.4 Temporal overview of the recordings

Since the recordings are rather intricate and time consuming, the experiments were run subject by subject over a period of more than two years, rather than by carrying out several experiments in parallel.

The speakers wore the palates for two weeks and were recorded several times during this period, three times via EMA and parallel acoustic recording

and twice via acoustics only. The first recording consisted of three subrecordings. The speakers were at first recorded with their natural palate (recording E1/1); afterwards the artificial palate was inserted. The auditory feedback was masked by white noise and the speaker spoke the complete corpus once again (recording E1/2). Care was taken that speakers did not speak with the prosthesis before the recording started. During the third subrecording auditory feedback was available and the speaker spoke the corpus a third time (recording E1/3).

The first acoustic recording (recording AU1) was carried out after the speakers had worn the prosthesis for three or four days. The subject spoke the complete corpus once with the artificial palate.

The second EMA recording (recording E2/1) took place after one week. The subject spoke the corpus once with the artificial palate.

The second acoustic recording (recording AU2) was carried out at day 11 or 12. The conditions were the same as for the first acoustic recording.

The third EMA recording took place on day 15. At first, the subject spoke the corpus with artificial palate (recording E3/1), afterwards the palate was taken out and the corpus was recorded once again without the artificial palate (recording E3/2). The following list summarises all the recordings and the different conditions.

- Day 1:
  - session E1/1: without artificial palate
  - session E1/2: with artificial palate, with auditory feedback masking due to white noise over headphones
  - session E1/3: with artificial palate with full auditory feedback
- Day 3 or 4:
  - session AU1: with artificial palate
- Day 8:
  - session E2/1: with artificial palate
- Day 11 or 12
  - session AU2: with artificial palate
- Day 15:
  - session E3/1: with artificial palate

– session E3/2: without artificial palate

The experimental setup, especially for the first session, is extremely complex and difficult to coordinate. The first recording with speaker TP took about 2.5 hours (around 1.25 h pure recording time plus preparation). With practice the recording time could be shortened a little. Since speakers tend not to tolerate recordings which are longer than two hours very well, session EMA1/2 was left out for two speakers (KD and OP) who were recorded as second and third speaker, respectively.

4.5 Auditory feedback masking

During the first perturbed session auditory feedback masking was carried out via the presentation of band pass filtered white noise (100 Hz-10 kHz). For the presentation of this white noise a system had to be found

- which would minimally interfere with the EMA system
- (for the recordings in Berlin) which would fit under the EMA helmet
- which would not emit noise into the recording room so that it is recorded by the microphone.

More specifically, this means that the system should not contain much metal; furthermore, the headphones could, at least for the recordings in Berlin, not be thicker than about 2cm; finally, closed headphones had to be used in order to prevent emission of the noise into the recording room.

For the recordings in Berlin a hose cable was used. The basis of the setup was a toy stethoscope made of plastic. The noise was carried from a CD player outside of the recording room to small ear phones. These ear phones were placed into the connecting part of the two hearers of the stethoscope.[3] The two hearers were put into the ears of the speakers. They were small enough so that they fitted in between the helmet and the subject's head. In order to prevent noise emission into the recording room the complete stethoscope was wrapped in pieces of carpet and isolation tape. Since the subjects experienced the stethoscope as extremely uncomfortable to wear the ends of the hearers were softened by attachment of Optosil and cotton.

The Munich partner used ear phones which were placed directly into the subject's ears. In order to prevent noise emission, an ear protection head set was placed on top and pushed against the subjects head with a rubber tape

[3]Thanks to Alan Wrench for raising this idea.

which was attached around the subjects head. The ear protection head set could be used because there is more space in the setup of the 3D EMA than in the one of the 2D EMA.

4.6 Corpus

Results from a pilot study carried out with one speaker and several prostheses types made of Optosil had shown that:

- fricatives are most difficult to adapt
- high vowels are more difficult than low vowels
- voiceless stops are more difficult than voiced stops
- bilabial stops are not perturbed very much by palatal prostheses

Furthermore, it was assumed that speakers would invest more effort in the adaptation of stressed tense vowels compared to other vowels. Consequently, the following target sounds were recorded: /z, s, ʃ, ç , x, t, d, k, g, i, e, y, o, u/.[4] They were embedded into nonsense words of the structure $C_1V_1C_2V_2$ where C_1 is one of the consonants and V_1 one of the vowels. There were three exceptions. /s, ç/ and /x/ were put in the C_2 position since they do not occur word initially in Standard German.

The remaining vowels in the nonsense word were always /a/, in order to have large tongue movements. There is one exception. Since /ç/ occurs only after front vowels in German, the preceding vowel was set to /ɪ/. In nonsense words with vowels as target sounds the consonants were always /t/ since this sound was assumed to be not very much influenced by coarticulation[5] so that the tongue would move towards the vowel from approximately the same place for all the vowels. It was decided not to take non-lingual consonant as for example /b/ because then the tongue movement for the vowel would have started during this consonant already and one would not have movements from a rather fixed point such as the alveolar ridge for all vowels. The resulting nonsense words can be found in table 4.2.

The carrier phrase was *Ich sah ... an.* (I looked at ...). The sentences were in general repeated 20 times per session in randomised order. This means that per subsession 220 sentences were recorded.

[4] In the present study the focus is laid on the investigation of fricative and vowel adaptation.

[5] See Daniloff et al. [1980] who show that coarticulation depends on the size of the phoneme inventory. For English they show that the variability of /t/ is minor. Since the stop inventories of English and German are very similar we assumed the same for German.

Table 4.2: Nonsense words used in the study. Target sounds are printed in bold

orthographic presentation on the screen	phonological transcription
sassa	/'zasa/
schacha	/'ʃaxa/
dicha	/'dɪ**ç**a/
dascha	/'daʃa/
katta	/'kata/
gacka	/'gaka/
toota	/'to**ː**ta/
tuhta	/'tu**ː**ta/
tühta	/'ty**ː**ta/
teeta	/'te**ː**ta/
tiehta	/'ti**ː**ta/

4.7 Acoustic analysis

This section describes the acoustic segmentation (section 4.7.1), the formant measurements carried out for the vowels (section 4.7.2) and the acoustic analysis of the fricatives (section 4.7.3) and a statistical method used in order to describe spectra 4.7.4.

4.7.1 Acoustic segmentation

The acoustic data were segmented into the speech sounds making up the nonsense words by students working at the two laboratories involved. For the segmentation PRAAT (Boersma and Weenink [1992–2004]) textgrids were used. The data of the first two speakers were segmented manually. Afterwards, a program for automatic segmentation set up by the Munich partner was used. This program enabled a semiautomatic segmentation. For each speaker one example sentence for each item was manually segmented. Afterwards, the program created textgrids for all the other repetitions of all the other sessions for the same item by using dynamic time warping, a procedure which compares two sequences and aligns them nonlinearly in the temporal dimension. These automatically generated textgrids were afterwards manually corrected.

With respect to the criteria of the segmentation, the onset and offset of the vowels were defined as onset and offset of the second formant. The onset of a stop was defined as the beginning of the silent interval. The offset of the stop was defined as the burst. Onset and offset of fricatives were defined

as beginning and end of the frication period. Consequently, offset of one sound and onset of the following sound usually did not coincide. After the segmentation had been carried out, a program provided by the IPS Munich was run which created one file per subsession with all the labelled points saved to a matrix.

4.7.2 Acoustic analysis of vowels

The speech material of this study makes automatic formant calculation rather complicated for several reasons. The perturbation leads to more variability so that fixed boundaries in the frequency domain within which one could look for formants could not be used. Also, the vowel as measured in the acoustic segmentation often did not show a very stable formant pattern over its complete duration. Furthermore, the vowels were, especially in the first perturbed session, sometimes partly or fully devoiced. Another problem was a high degree of friction or even complete contact during the production. A completely automatic method where boundaries of phoneme regions were defined once for each speaker and each vowel, which had been used for a previous study turned out not to give satisfactory results for the present corpus. On the other hand, the size of the corpus did not allow for a completely manual strategy. Formants were therefore measured by a semiautomatic method set up by the IPS Munich and the ICP/Gipsa-lab Grenoble. This program works with two methods, peak picking and root solving, and takes into account a specification of upper and lower boundaries of formants in the spectral domain.

After the calculation of the FFT and an LPC with 14 coefficients for male speakers and 12 for female speakers with Matlab functions the program calculated a preliminary estimate of the formants with a root solving method. The calculations of formants $\hat{F}1$, $\hat{F}2$ and $\hat{F}3$ were carried out over the duration of the complete trial with a window length of 256 samples and 50% overlap of windows. The results for each trial were saved to a file.

Then these trial files were loaded one after the other in the analysis program *mt_new* (Hoole [2007]) in order to carry out the semiautomatic measurements. A spectrally stable segment of the vowel was chosen by adjusting the cursors around it in the temporal domain. Then the measurement program was started. Boundaries for each peak were specified, at first automatically, but they could be adjusted manually later.

Then the actual calculation of formants started. The window duration was set to 20ms with an overlap of windows of 5ms. The data were resampled to 10 kHz and a preemphasis of 6dB/octave was carried out. A hanning window was used. At first, the roots of the LPC were calculated. A fixed

LPC order of 14 was used. Within each set of boundaries the resonance with the smallest bandwidth was taken as the formant. Then the same measurement was carried out with the peak picking method. The value with the highest amplitude within the segment was taken as formant. If there was no maximum, a region with constant values was looked for by searching for a minimum of the derivative of the LPC-curve.

The values which were found for the two methods were plotted together with the LPC of the segment as asterisks at the spectral peaks. If they did not represent the formants because, for example, two formants had been merged, the formant boundaries were adjusted manually and the process was repeated.

In a last step the results of the two methods were compared. In the rare case that they differed a lot the more untypical value (compared to the other repetitions of the same session) was removed.

4.7.3 Acoustic analysis of obstruents

In contrast to the description of vowel spectra which can easily be done by formant measurements, the description of obstruent spectra has until now been under debate.

The majority of descriptions uses methods from statistics originally designed to describe probability functions, e.g. the calculation of moments. The moment of the order k in relation to r is calculated as:

$$m_k(r, x) = \int_{-\infty}^{\infty} (x - r)^k * p(x) dx \tag{4.3}$$

where r is the centre of the probability distribution, k is the number of the moment and $p(x)$ is the probability distribution itself.

The most commonly used moment for the description of obstruent spectra which is used especially in phonetics and phonology is the first moment, the mean value of the spectrum, also called *centre of gravity*. This parameter enables to draw very broad distinctions between phonemes based on the place of articulation and the resulting filtering of the noise after having passed the constriction. For example, during a velar fricative the noise passes a rather long tube from the constriction to the lips so that the amplitudes of rather low frequencies are raised. In contrast to that, during an alveolar fricative the tube leading from the constriction to the lips is short which results in higher amplitudes of high frequencies. In the first case the centre of gravity is therefore lower than in the second.

Even if this parameter is widely used it does not succeed in drawing clear distinctions between phonemes. It is even less useful in cases where finer

grained differences are investigated as for example in perturbed or pathological speech.

The second moment (with $k = 2$ in equation 4.3), called dispersion is in statistics the standard deviation. In relation to spectral characteristics this measurement is an expression of the compactness of the spectrum. If the spectrum is peaked (as for example in /s/), the value is low. If there are no prominent peaks, as in labial fricatives, it is high. This measurement succeeds rather well in drawing the distinction between flat and peaked spectra. However, on its own it does not succeed in differentiating between phonemes.

Forrest et al. [1988] calculated the third and the fourth moment and from these values *skewness* and *kurtosis* of fricative and stop burst spectra. Skewness was defined as the third moment divided by the second moment (the standard deviation) to the power of $\frac{3}{2}$:

$$\nu = \frac{m_3}{\sigma^{\frac{3}{2}}} \tag{4.4}$$

This value gives information about spectral asymmetry. If the spectrum is skewed to the right the value is below 0, if it is skewed to the left the value is greater than 0. High absolute values denote a high degree of skewness. To give an example, the spectrum of the alveolar fricative /s/ has a high negative skewness whereas the spectrum of /x/ has a high positive skewness.

Kurtosis is the fourth moment normalised by the second power of the standard deviation:

$$\beta = [\frac{m_4}{\sigma^2}] - 3 \tag{4.5}$$

It can be interpreted as the peakedness of the spectrum. Spectra with one prominent peak as for example the alveolar fricative have a high kurtosis.

Evers et al. [1998] developed a further method in order to characterise spectra and thereby distinguish between the alveolar and the postalveolar stop. They calculated a regression line over the spectrum from 0 to 2.5 kHz (called *a-slope*), and another one from 2.5 kHz to 8 kHz (called *b-slope*). For the alveolar fricative both slopes where positive whereas for the postalveolar stop the a-slope was positive and the b-slope was either positive, but much smaller than the one for the alveolar stop, or it was negative.

Beginning and end of the fricatives were labeled as described in section 4.7.1. For the calculation of these parameters, the signal was at first high pass filtered (700 Hz) in order to remove influences of voicing. The upper cut-off frequency (12 kHz, resulting from the filtering during preprocessing) was chosen rather high because preliminary investigation of the spectra showed that perturbed and unperturbed spectra differed considerably in the higher frequency ranges. The analysis window (Hanning) for the calculation was 8

ms and was moved by 5 ms increments resulting in 60% overlap. An FFT was computed for each analysis window. Afterwards, a power spectrum was calculated which was then normalised so that it shared the properties of a probability density function (i.e. the area below the curve is 1).

The spectral moments, a-slope and b-slope, were calculated for this function. For the slope measurements a linear approximation was carried out for the power spectrum from 700 Hz to 2.5 kHz and from 2.5 kHz to 12 kHz. The slope of the resulting two functions was calculated. The actual calculation of all these parameters was carried out with a program provided by the ICP/Gipsa-lab Grenoble. Afterwards a mean for each parameter was calculated over all the windows of each sound.

4.7.4 Discriminant analysis

The discussion has so far shown that for the description of both vowel and fricative spectra more than one acoustic parameter is necessary since each of the parameters sheds light on a different aspect of the fricative or vowel spectrum. For the further analysis, discriminant analyses with all these parameters entered as variables were calculated.

A discriminant analysis is a multivariate statistical procedure which can be used in order to investigate in how far a dependent variable out of a group of dependent variables is useful in order to classify observations. Very generally, this is done by calculating a discriminant function as a linear combination from the variables which enter the analysis, whereby the coefficients of the linear combination are chosen in a way to maximise the difference between the groups. The resulting discriminant function can be used in order to assign each observation to the group to which it fits best judging from its factor values (resulting from the discriminant function). This is done by calculating the difference between the factor values and the means of factor values of the observed groups. From these differences the probability of a production to belong to a group is calculated (see Bortz [1999] for more details on discriminant analysis and classification procedures).

For the data discussed here, the discriminant analysis will be used for two kinds of investigations: For a phoneme classification of vowels and fricatives (two separate analyses) and for a judgment of the quality of the fricative productions. The calculations were carried out in SPSS 15.0. In the first case (phoneme classifications) the discriminant analysis is used in order to assign each fricative or vowel production to the phoneme class to which it fits best according to its acoustic properties. If the classification of the productions correlates highly with the observed group membership (the productions are classified as the phonemes which they were intended to represent) this

suggests that (1) the phonemes can be classified according to acoustic parameters and (2) the acoustic parameters which were measured represent well the parameters which are necessary in order to draw the distinctions among the phonemes. If 100% of the productions are classified as the phoneme which they were intended to represent this suggests that all the information which is necessary in order to classify the sounds can be found in the acoustic domain.

All the measured parameters (COG, dispersion, skewness, kurtosis, a-slope and b-slope for the fricative analysis and the formants and duration for the vowel analysis) entered the discriminant analysis as dependent variables. The phoneme was the group variable. Then the discriminant function was calculated as a linear combination of the six variables (for fricatives) and four variables (for vowels) whereby the coefficients were chosen in a way to maximise the difference among the phonemes. All parameters were included at once in order not to lose interaction effects among parameters. Afterwards, the probability of each production to belong to a phoneme class was calculated.

For this first analysis data were split by speaker. After an initial analysis without split by session had shown that the productions could not well be classified it was decided to split the data by session as well. For a discussion of this see chapter 5.

The second discriminant analysis had the aim to assess the quality of the fricative productions in order to see a possible improvement over time. Data were split by speaker and phoneme and a discriminant function was calculated for the data from sessions E1/1 and E1/2 which maximally separated these two sessions. The six acoustic parameters were entered all at once as dependent variables and the session was taken as group variable. For the classification, all data from all sessions were included and the probability to belong to sessions E1/1 or E1/2 was calculated.

When a production shares many acoustic characteristics with the "typical" unperturbed production (it is thus a "good" production) the probability to belong to session E1/1 is high. Conversely, when a production shares many acoustic characteristics with the initial perturbed productions where speakers had hardly had a chance to practice (it is thus a "bad" production) the probability to belong to session E1/1 is low. The expectation was that in the early perturbed sessions there would be very many "bad" productions which are very much like the productions in session E1/2 and only afterwards the productions would improve and become more like the "good" productions in session E1/1.

4.8 Articulatory analysis

Two kinds of analyses were carried out on the articulatory data. Articulatory target positions for the sounds were measured (section 4.8.1). Afterwards, articulatory gestures were segmented (section 4.8.2) in order to analyse kinematic parameters.

4.8.1 Articulatory positions

For each vowel and each consonant production an articulatory target position was measured. Very generally, this is the point in time plus the corresponding positional data when the tongue has reached a certain extreme position (usually a very high position for consonants and a lower position for vowels) and the movement direction changes. In order to measure these positions, a definition of each sound target was set up. These definitions can be found in table 4.3. For vowels the target is usually defined as the highest point (the highest y-value) of some tongue sensor during the acoustically defined interval $C_1V_1C_2$. Exceptions are /u/ and /o/, where the most posterior position (the highest x-value) was selected. The position of the consonants was searched for within the acoustically defined interval VC_1V_1 or $V_1C_2V_2$. Normally, the targets of alveolar and postalveolar consonants were defined as the highest position of the tongue tip sensor; the targets of the other consonants were defined as the highest position of the tongue back sensor. The selection of the criteria for the targets was driven by two considerations, first, the involvement of the articulator (tongue tip, tongue back) in the formation of the sound and second, and more practically, the reliability and consistency of a criterion for the measurements. The definition for /s/ is driven by the first consideration, because the tongue tip is actively involved in the formation of the constriction. The definition for /u/ is rather driven by the second consideration: The most retracted position of the tongue could reliably be identified on the trajectory.

The measurement of these points was carried out semi-automatically: At first the positions were determined automatically, afterwards they were manually corrected. The points in time of the articulatory target and the positions of the respective sensor were saved to a file.

4.8.2 Articulatory segmentation

Articulatory segmentation has the aim to find articulatory gestures. A gesture is in this case defined as a movement of an articulator from the configuration of one sound to the configuration of another one. In the nonsense

Table 4.3: Parameters defining articulatory targets. Left column: Sounds for which target measurements were carried out. Right: definition of the articulatory target.

sound	parameter
/a/	lowest y-value of tongue back coil
/o/	highest x-value of tongue back coil
/u/	highest x-value of tongue back coil
/e/	highest y-value of tongue blade coil
/i/	highest y-value of tongue blade coil
/ɪ/	highest y-value of tongue blade coil
/k/	highest y-value of tongue back coil
/g/	highest y-value of tongue back coil
/t/	highest y-value of tongue tip coil
/d/	highest y-value of tongue tip coil
/ç/	highest y-value of tongue back coil
/x/	highest y-value of tongue back coil
/ʃ/	highest y-value of tongue tip coil
/s/	highest y-value of tongue tip coil
/z/	highest y-value of tongue tip coil

word /'zasa/, for example, one can find four tongue gestures, one from the preceding /a/ of the carrier phrase to the voiced alveolar fricative, one from this fricative to the /a/ following it, one from this vowel to the voiceless fricative and one from the voiceless fricative to the final /a/.

There are consonantal gestures (leading towards a consonantal configuration, for example the gestures towards /z/) and vocalic gestures (leading towards a vocalic configuration). Looking at the vertical movement of the articulators, consonantal gestures normally involve upward movement whereas vocalic gestures involve downward movement.

For the gestures segmented here a method proposed in Hoole [1996b] was used: Since onset and offset of a gesture are not always easy to determine on the positional signal, they were segmented on the velocity signal as the segment which could be determined by two velocity minima around a maximum (cf. figure 4.3 showing the segmentation of a vocalic and a consonantal jaw gesture). In the majority of cases the maximum could easily be found. Due to movements along the palate (as for example in loops[6]) or intervals where the articulators just do not change position (e.g. stop closure), there are often multiple small peaks and minima in between of two big maxima.

[6]sliding movements along the palate which occur mainly in velar stops

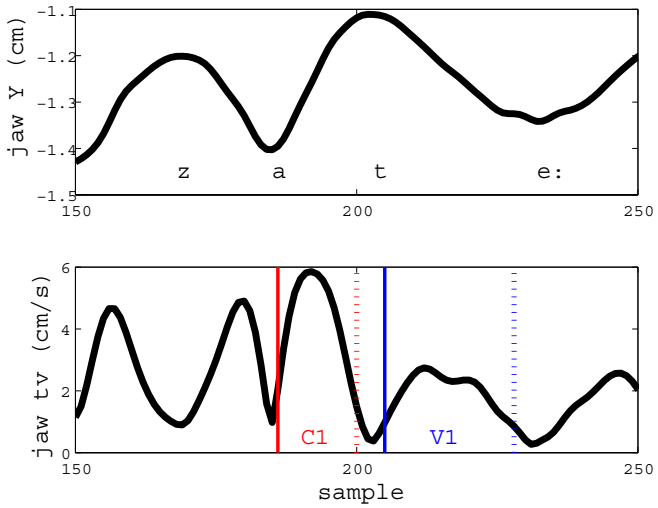


Figure 4.2: Example for measurements of articulatory gestures in the sequence /za'teː/ (*sah* from the carrier phrase and beginning of test word *teeta*). Upper subplot: Vertical movement of the jaw sensor. Lower subplot: Corresponding tangential velocity. The first pair of vertical lines marks beginning (solid) and end (dotted) of the consonantal gesture leading towards C1, the second pair of vertical lines marks beginning (solid) and end (dotted) of the vocalic gesture leading towards V1. Beginning and end of each gesture are 20% above a velocity minimum.

In order to achieve consistent measurements which are independent of these small minima a 20% threshold criterion was used: The amplitude from the first minimum to the maximal value was measured, 20% of the complete amplitude was calculated, and the point on the velocity signal which was 20% above the minimum was taken as onset of the gesture. The offset was defined as the latest point 20% above the amplitude value of the last minimum.

As described in section 4.1, velocities and accelerations for the x and y movement and tangential velocity and acceleration were calculated during preprocessing. Since the majority of gestures involves movements in vertical as well as horizontal dimension at the same time, the gestures of tongue and jaw were calculated on the tangential velocity signal.

As has often been discussed, the tongue and lower lip movement is dependent on the jaw movement (e.g. Westbury et al. [2002]). Several methods have been developed in order to separate the lip or tongue movement from the jaw movement. The problem for the calculation of this "intrinsic" tongue movement is that the jaw movement involves at the same time a translational

and a rotational movement. One can therefore not simply subtract the jaw movement from the tongue movement in order to get the intrinsic tongue movement. For the tongue tip the error resulting from assuming only a translational movement should be rather small, for the tongue back, however, it will be notable. Mooshammer et al. [2006] were able to calculate a rotational component by determining the position of the jaw joint on MRI recordings of the speaker and thus calculating the jaw angle. Another possibility to calculate intrinsic movements is to determine the jaw angle on the basis of a second jaw sensor placed closer to the jaw joint during the EMA recording. In our experimental setup a second sensor could not be included so that separating the jaw from the rest of the structure is not easily possible. Furthermore, we did not have information about the position of the jaw angle of our speakers from another source (e.g. MRI). Thus, the intrinsic tongue movement was not measured and complex gestures were used for further analysis.

Measurements were carried out in *mt_new* for all the gestures in between the preceding /a/ (from the carrier phrase) and the second vowel of the nonsense word. The gestures of the jaw, tongue dorsum and tongue back were segmented for all the sounds. Tongue tip gestures were segmented for all sounds except the velar stops and the velar fricative where in general no tongue tip gesture existed.

The measurements were normally unproblematic since most consonants followed /a/ and most vowels followed /t/. Consequently, the movements were rather large which resulted in clear maxima and minima in the velocity signal. The point of departure for the measurement was always the movement signal (y). For vocalic gestures of all sensors the velocity minimum matching closest with the highest y-value was taken as departure for the measurement of the onset and the velocity minimum matching closest with the lowest y-value was taken for the measurement of the offset. For consonantal gestures it was the other way around, for the onset measurement the velocity minimum matching with the lowest y-value was taken, for the offset the one matching with the highest y-value.

Some problematic cases have to be mentioned, however. During the production of /t/ the tongue often kept a constant y-value over a longer period which resulted in a low tangential velocity with several peaks and minima. In this case the point of departure was the middle of the closure time and the gestural onset was measured 20% above the minimum closest to it.

Especially in the first perturbed session the velocity profiles of the gestures were not unimodal but sometimes bimodal. In this case the gesture was measured as including both peaks, the maximum was measured as the higher peak of the two.

Articulatorily very similar sounds sometimes shared a gesture. In /'dɪça/ for example, there was usually only one tongue gesture for /ɪç/. In this case this one tongue gesture was taken as the measurement for both gestures, the vocalic and the consonantal gesture.

Chapter 5

Acoustic characteristics of vowels

The description of the experiment in chapter 4 will now be followed by several analytically oriented chapters. This series start with a rather fundamental analysis in speech, i.e. with the one of vowel acoustics. Vowels are traditionally described in terms of formant structure and duration. However, this description is limited by two factors. First, the formant structure depends not only on the vowel but also on the vocal tract size of the speaker so that a classification of vowels by formant structure is speaker dependent (e.g. Peterson and Barney [1952]). From a perceptual point of view, listeners thus have to take into account speaker specific characteristics in their processing in order to be able to classify vowel productions (cf. e.g. Miller [1989], Johnson [1989], Nearey [1989], Strange et al. [1976] for speaker normalisation and its problems). Second, the duration cannot be seen in absolute terms either, because it depends on the speech rate (cf. Tuller et al. [1982], Tuller et al. [1983], Tuller and Kelso [1984]), and a normalisation with respect to this rate has to be carried out by the listener.

Even when these two factors, speaker dependence and speech rate dependence, are taken into account, several productions of a phoneme do not show an invariant pattern, but the variability in acoustics is rather large so that the productions could be described as forming a region in acoustic space created by the formants, the duration and maybe other parameters. As mentioned already, Miller [1989] found that regions in formant space do not overlap for American English vowels. Assuming the same for German vowels, it should be possible for the listener to assign each production to a phoneme.

The analysis presented in this chapter has two aims. The first aim is to see whether separate regions in acoustic space for the vowels can be found in all sessions and the second aim is to investigate the development of the

formant values over time. The second aim, to describe the development, will not be accomplished in a satisfactory way in this chapter and a different analysis with the same aim will therefore be at the centre of the following chapter where articulatory data will be considered as well.

No matter what the phoneme representations are, whether acoustic or articulatory, given that the context by which the vowels are surrounded is the same in our sample, one should be able to separate the vowels acoustically. This is because speakers will need acoustic differences, either because they use the acoustic information directly or because they use the acoustic signal in order to extract the articulatory information which is necessary for the identification of the sound. This assumption, that separate regions can be expected even when speakers adapt towards articulatory targets seems to contradict one of the classical example supporting articulatory perceptual primitives, namely that the perception of the same acoustic signal (in this case a stop burst) differs with respect to the context (Cooper et al. [1952]). However, since the context of the vowels discussed here is the same (except for the lax vowel /ɪ/) separate regions in acoustic space should result, or at least the occurrence of separate regions should in general not disprove articulatory perceptual primitives. Thus, analysing regions in acoustic space as such does not contribute much to the central question of this study, however, as will be seen, the analysis will give important hints for further analyses.

Under perturbation speakers will over time produce vowels which can be separated acoustically. For our data informal perception tests suggest that all the vowels, even in the first perturbed session without auditory feedback can be correctly classified. It can therefore be assumed that the productions are all in the respective phoneme region in perceptual space. The first part of the analysis presented here thus investigates whether this perceptual result can be supported by acoustic measurements.

More informative for the basic question about the speech production tasks should be the second part of the analysis which deals with the development of the formant values and vowel durations over the adaptation time. When the aim of the adaptation is in the acoustic domain speakers can be expected to show a development towards values which are more and more like the ones of the unperturbed session. When the aim of their adaptation is in the articulatory domain, the development of the formant values might not be directed towards the original values, but there might be a development which has a clear equivalent in articulatory space. For example, there could be a rise in F2 suggesting more and more forwarded productions.

The structure of the chapter follows the analysis steps. At first single formants are discussed and an analysis of variance is calculated (section 5.1). Since, however, in order to classify each production it is not sufficient to

have significant differences but rather one needs to have separate regions in acoustic space, for each vowel pair the degree of overlap of the formants is investigated. In section 5.2 F2-F1 and F2-F3 formant space is discussed. Durations are analysed as a fourth parameter (section 5.3). In order to see whether the durational characteristics make up for a missing difference in formant values, trade offs between durations and formant values are investigated and a discriminant analysis is calculated in order to find out whether it is possible to classify the productions when all four parameters (F1, F2, F3, duration) are taken into account (section 5.4).

5.1 Single formants

Formants were measured as described in section 4.7.2. Afterwards, the influence of the vowel identity on each formant was statistically tested. Data were split by speaker and ANOVAs were calculated with *vowel* and *session* as factors and the formants as variables. Unsurprisingly, a significant influence of the vowel identity on each formant was found (results are given in the appendix, table A.1, p.142). However, the influence of the *session* and the interaction *vowel*session* were also significant.

Tamhane T2 post-hoc tests[1] (data split by speaker) showed that not all the differences between vowel pairs are significant. Again unsurprisingly, vowels with similar tongue heights often showed no significant difference in F1, vowels with similar anterior-posterior position often had no significant difference in F2. For example, the difference in F1 between /e/ and /ɪ/ was not significant for three speakers, and the difference in F2 was not significant for /o/ vs. /u/ for two speakers. This is shown in table 5.1 which gives the cases where no significant difference between two vowel categories was found in the post-hoc test. The results show that there is no vowel pair which is not distinguished by at least one formant.

Even if the formant structure differs significantly for each vowel pair, a significant difference is not sufficient since it allows for some overlap between phonemes. In order to quantitatively assess the degree of overlap, means and standard deviations for each formant value (split by speaker) were calculated. Formant "regions" for each vowel (data split by speaker) were defined by mean values ± 2 standard deviations. The overlap of these regions for each vowel pair was investigated. Tables A.2 to A.5 in the appendix (p.143ff) give the results. In cases where there is overlap in all three formants the numbers are printed in bold. As one can see, the overlap is considerable

[1]Throughout the study Tamhane T2 post-hoc tests were calculated rather than for example Scheffé tests because the error variances differed for the sessions.

Table 5.1: Cases where formant values did not differ significantly for a vowel pair according to Tamhane T2 post-hoc tests.

formant	vowel pair	speakers
F1	/e/ vs. /ɪ/	AP, DS, TP
	/i/ vs. /y/	AP
	/o/ vs. /ɪ/	AP
	/i/ vs. /u/	BP, DS
	/i/ vs. /y/	BP
	/u/ vs. /y/	BP, DS, OP
	/o/ vs. /e/	DS, KD, SK
	/ɪ/ vs. /e/	DS
F2	/e/ vs. /i/	AP
	/o/ vs. /u/	KD, SK
	/i/ vs. /ɪ/	TP
F3	/e/ vs. /i/	AP, DS
	/e/ vs. /ɪ/	AP, DS, TP
	/ɪ/ vs. /i/	AP, BP, DS
	/o/ vs. /e/	BP, DS
	/o/ vs. /u/	KD

and the number of vowel pairs where there is overlap in all three formants is remarkable as well.

5.2 Overlap in F2-F1 and F2-F3 space

One has to consider, however, that overlap in all formant values does not necessarily mean that there are no separate regions in formant space. For example, two F2-F1 ellipses with a main orientation of 45°, which have the same x-values but one of them having higher y-values, might not overlap although there is overlap in both x and y-values. That is why F2-F1 and F2-F3 space will be investigated. For reasons of illustration this method will at first be discussed with the help of two examples in 2D-space, the first one in F2-F1, the second one in F2-F3 space.

Dispersion ellipses (2 standard deviations) were calculated for data split by speaker and inspected visually for overlap between phonemes. At first, all productions of all sessions were plotted together. An example for the F2-F1 space is given in figure 5.1[2]. The figure shows the F1 and F2 measurements

[2]The low vowel /a/ is plotted for reasons of illustration but it will not be further dis-

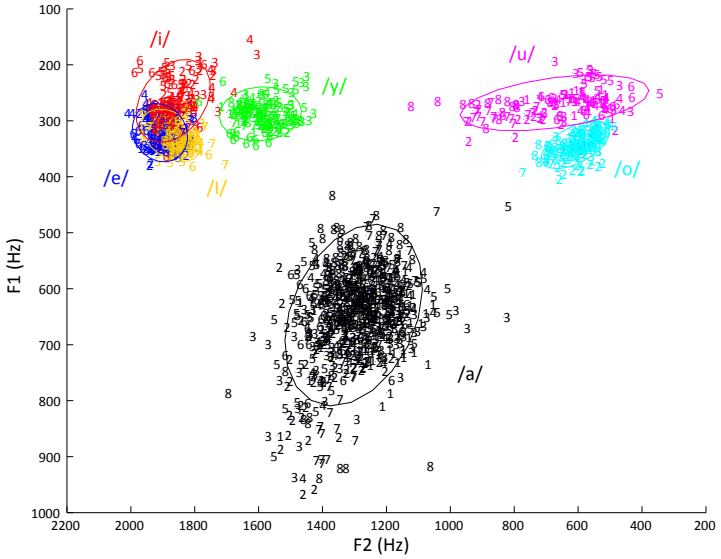


Figure 5.1: F2 (abscissa) and F1 (ordinate) of the vowels produced by speaker TP. Grey shades denote different sounds, the numbers denote the sessions: 1: E1/1, 2: E1/2, 3: E1/3, 4: AU1, 5: E2/1, 6: AU2, 7: E3/1, 8: E3/2.

of the productions by speaker TP. The different grey shades denote different vowels. The numbers give the session (in chronological order) in which the sound was produced. The figure shows very clearly that for each vowel there are preferred regions which are rather well separated, except for the three front vowels for which there is some overlap. Consequently, even if the formant values differ significantly for the front vowels, the two dimensional plots show that they do not differ enough in order to unambiguously assign each production to a phoneme.

The case of speaker TP shown here is typical for the whole group of speakers if one looks at productions pooled over sessions: Normally there are a number of vowels which do not have separate regions in F2-F1 space, namely /i, ɪ/ and /e/, and for some speakers also /o/ and /u/, for DS also /y/ and /i, ɪ, e/. Inspection of F2-F3 plots revealed that for speaker DS, where the difference between /y/ and other front vowels was not clear in F2-F1 space, it is not clear by F3 either, even if the centres of the ellipses are further apart in F2-F3 space than in F2-F1 space.

Looking at figure 5.1 more closely, one can see that the productions of a session can be found in a preferred subregion of the region created by

cussed since it was easy to adapt and is generally very variable articulatorily and acoustically without overlap with the other vowels investigated.

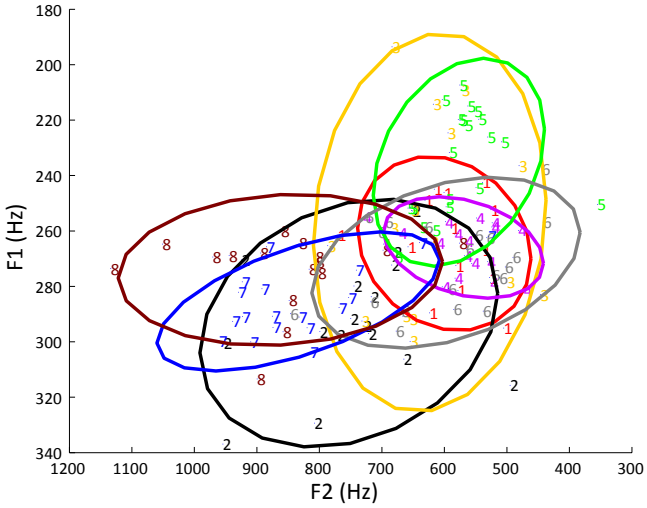


Figure 5.2: F2 (abscissa) and F1 (ordinate) values of /u/ for speaker TP. Grey shades and numbers denote different sessions in chronological order (cf. figure 5.1). Ellipses (2 standard deviations) show the subregions of the vowel.

all productions of a phoneme over sessions. In other words, the regions in formant space vary more across sessions than within sessions. In order to illustrate this point figure 5.2 shows the data for /u/ presented in figure 5.1, but enlarged and with different grey shades for different sessions. The region which is created by the ellipses of all productions across sessions will from now on be called *region*. The region for one recording only will be called *subregion*.

Each session thus occupies a certain subregion within the region in acoustic space. Session 1 (E1/1) is rather in the centre of the region, session 5 (E2/1) is in the upper part of the figure, sessions 7 and 8 (E3/1 and E3/2) are in the left part of the F2-F1 space. Informal perception tests suggest that all these productions are good representatives of their phoneme class. Furthermore, one cannot detect a clear development in the pattern. It is not the case that, e.g. F1 becomes higher and higher over time or that session 2 is furthest away from session 1 and afterwards there is a development towards the initial values. The inconsistent development which can be seen in this figure is very characteristic for all speakers and all vowels.

To give a second example, as mentioned already, the distinction /y/ versus /i, e/ and /ɪ/ for speaker DS in F2-F1 and F2-F3 space was not clear. The problem here is that F3, which could create the difference between the three unrounded vowels and the rounded vowel, is rather variable. A closer

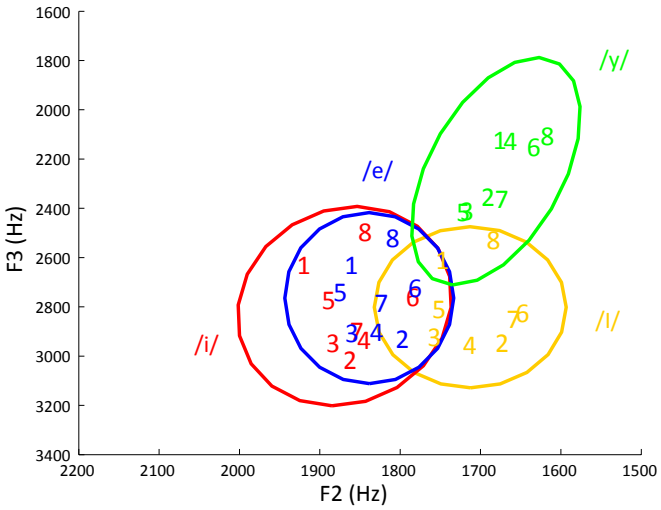


Figure 5.3: F2-F3 space for the productions of /i, e, ɪ/ and /y/ of speaker DS. Numbers give means of sessions in chronological order (cf. figure 5.1), grey shaded ellipses (2 standard deviations) give the regions corresponding to a phoneme.

inspection of the different sessions, however, reveals that for each session separately, there is in most cases no overlap in F3. Figure 5.3 is meant to illustrate this. One can see the *regions* (over all sessions) of each front vowel as ellipses in different grey shades. The numbers show the centres of the *subregions* (one number for each session in chronological order). Looking at session 1 (E1/1) F3 is rather low[3] for /y/ ("1" within the ellipsis in the upper right of the figure), but it is also lower for the three unrounded vowels (other "1s"). If one compares session 1 to session 2 (E1/2) one can see that in this session F3 is in general a bit higher for the rounded but also for the unrounded vowel. Thus, the productions of each subsession are in a certain subregion within the phoneme region measured over sessions. Although there is overlap in the *region* there is none in the *subregion*.

This result which was shown for the rounded-unrounded front vowel contrast can be found for the other tense vowel contrast as well, i.e. the /i-e/ contrast. Dispersion ellipses thus do not overlap as much in both F2-F1 and F2-F3 space when only the productions within one session are analysed. Formant regions therefore cannot be seen in absolute terms, but they exist in relative terms for each session.

Table 5.2 gives the cases in which there is overlap in both the F1-F2 and

[3]Note that the axes are reversed.

Table 5.2: Cases in which there was overlap in both F2-F1 and F2-F3 space according to visual inspection of plots for data split by speaker and session. Column 1: speaker, columns 2-5: vowel pairs for which there is overlap. A "+" means that there is no overlap in duration and the sounds can thus be distinguished by this further characteristic (see section 5.3 for details.)

speaker	E1/1	E1/2	E1/3	AU1
DS	ei, eɪ, ou	eɪ	ei, yɪ	ei+, eɪ
KD	eɪ		eɪ	eɪ
OP	eɪ		eɪ, ei+, iɪ	eɪ
SK	eɪ, ou+	eɪ, ou	eɪ, ou	ei, eɪ, ou
TP	ɪi, eɪ	iɪ, ei+	ei, eɪ, ɪi	ei, eɪ, ɪi
BP	iɪ	iɪ	iɪ	eɪ, iɪ, ie
AP	eɪ, ou	eɪ, ie, iɪ, ou	iɪ, ie, eɪ, ou	eɪ
speaker	E2/1	AU2	E3/1	E3/2
DS	ei+, ɪi, eɪ	ei+	eɪ	eɪ
KD	eɪ		eɪ	eɪ
OP	eɪ, ie+		eɪ	eɪ
SK	eɪ, ou	ei+, eɪ, ou, ɪi	eɪ, ou+	eɪ, ou
TP	eɪ	ei, eɪ, ɪi	eɪ, ei+, ɪi	eɪ, ei+
BP	ɪi	iɪ	iɪ, ɪe, ei+	iɪ, eɪ
AP	ei+, iɪ, eɪ, ou+	ei+, iɪ, eɪ, ou+	ie+, eɪ, ou	ie+, iɪ, ou+

the F2-F3 space even within a session. Most cases involve the lax vowel /ɪ/. Apart from the overlap for this one vowel there is sometimes overlap between /i/ and /e/, and for speakers SK and AP for /o/ and /u/.

5.3 Vowel durations

Since it has turned out that some vowel pairs are not clearly distinguished in formant space (there is some overlap even if one splits data by session), this section will look at the duration of the vowels. Durations were calculated as the time from F2 onset to F2 offset, as described in section 4.7.1. In order to account for differences in speech rate the percentage duration with respect to the word length was calculated.[4]

[4]This of course presupposes that the words in which the sounds were uttered did not generally differ in length. Since the structure of the words was for all tense vowels /tVta/ it is assumed that this is the case. An exception is /ɪ/ which was uttered in the surroundings /dVça/. For this vowel it is therefore possible that the length measurement is slightly influenced by word duration.

An ANOVA with *vowel* and *session* as factor and *duration* as dependent variable for data split by speaker showed that a significant main effect for *vowel* and *session* was there in all cases. For *vowel*session* there was a significant main effect for all but one case (speaker BP, cf. table A.6, p.147 in the appendix). In general, the high vowels /i/, /y/ and /u/ tended to be shorter than the intermediately high vowels /e/ and /o/. Tamhane T2 post-hoc tests showed that the differences among /i, y, u/ and the ones between /o, u/ on the other hand, which were problematic in formant space are always significant except for the /o/-/u/ contrast of speaker BP. /ɪ/ is often not the shortest vowel and the difference to the other front vowels is often not significant.

Again, ellipses with 2 standard deviations were visually inspected for *duration-F1* and *duration-F2* space. The lax vowel belonged to the shorter vowels, but for separating it from the tense vowels the differentiation by formant values was generally better than the one by duration. Therefore, the duration of the lax vowel was not further investigated. For the tense vowels, a number of vowel distinctions became clear by a separation in *F1-duration* space. These are marked with a "+" in table 5.2. The *F2-duration* plots did not further contrast the vowel pairs.

Thus, even when each session is analysed separately, the overlap in formant values of the lax vowel /ɪ/ with the other two tense front vowels is substantial and the duration does not make clear the distinction either. This suggests that there must be a further mechanism which distinguishes tense and lax vowels. One possibility for such a characteristic, which has so far not been investigated for German but for Australian English, is the formant trajectory from the preceding sound over the vowel to the following sound. As shown in Watson and Harrington [1999] the target on this trajectory is reached later for the tense vowel than for the lax vowel. So it could be that speakers use the information from time varying differences in formants rather than the targets only which were measured here.

Another possible explanation could be that the lax vowels are actually not phonemic in German, as proposed by Vennemann [1991]. In this approach tense and lax vowels are characterised by differences in the way the syllables are cut, i.e. by differences in the energy contour over the syllable. The so far called tense vowels would occur in syllables where the *crescendo* (rising energy) and the *decrescendo* (falling energy) take place during the vowel mora. These syllables are called *smoothly cut syllables* ("sanft geschnittene Silben"). The lax vowels occur in syllables where only the crescendo is at the same mora as the vowel. These syllables are called *abruptly cut syllables* ("scharf geschnittene Silben"). Since every syllable needs to have a crescendo, an abruptly cut syllable cannot be open. Related to the material here, in

order to make the contrast between /i/ and /ɪ/, rather than using acoustic information such as the formant values, it could be sufficient for speakers if they recognise that the first syllable in /dɪça/ is abruptly cut (there is a crescendo during /ɪ/ and /ç/ occupies the coda with the decrescendo), whereas the first syllable in /tiːta/ is smoothly cut. Different regions in formant space for tense and lax vowels would thus not be necessary.

Yet a third explanation for how tense and lax vowels are distinguished acoustically would be that f0 is used as a further acoustic cue (Hoole and Mooshammer [2002], Hoole [2006]). The fundamental frequency is about the same in a tense-lax vowel pair, within the group of tense vowels or within the group of lax vowels, however, f0 is higher for high vowels. Sounds sharing F1 and F2 values but not belonging to the same tense-lax pair, as for example /e/ and /ɪ/, can thus be distinguished by f0.

5.4 Further investigation of the relation between formant values and durations

From a perceptual perspective, one can assume that in production the length characteristic might only be used if the formant pattern is not sufficient in order to identify the vowel. For example, as long as the F1 of an /e/-production is high enough in order to distinguish it from /i/ and low enough in order to distinguish it from /ɛ/ (which is not investigated here but phonemic in German) it can be short. If the F1 deviates from the "ideal", however, the vowel should become longer. Even if this has not been investigated systematically so far, Ziegler and Hoole [1989] report such a trade off for /i/ and /ɪ/ productions one speaker (cf. their figure 1 speaker N, p.454). For their consecutive perception experiment productions by a normal, two aphasic and a dysarthric patient were presented to listeners. In one condition the durational cue was removed from the stimuli. The results show that as long as the formant patterns differed sufficiently the vowels were correctly classified. As soon as the difference was too small, however, listeners needed a durational cue in order to correctly classify the sounds.

For the productions by the speakers in the present study the existence of this trade off was investigated with the following hypotheses:

- (1) For /ɪ/: If a production of this vowel has a high F2, it could be confused with /e/ or /i/. Since these vowels are long, shortening the /ɪ/ should be useful for its identification. Consequently, one can expect a negative correlation between F2 and duration for /ɪ/.

- (2) For /e/ and /o/: If the vowel height is not chosen correctly and F1 deviates from the "ideal" these intermediately high sounds could be confused with the high sounds /i/ and /u/ or with the lower ones /ɛ/ and /ɔ/. Prolonging the sounds should help to identify them. Consequently, there should be a correlation between the duration and the F1 deviation from the mean of the unperturbed session in /e/ and /o/.

In our data this trade off could not be found. The trade off between duration and formant values thus seems to be predominantly perceptual and is not used in order to facilitate adaptation.

In order to investigate the combined influence of all four characteristics, a discriminant analysis (cf. section 4.7.4) was carried out. Data were split by speaker and, since it has turned out that formant and duration values vary over sessions also by *session*. For the discriminant function *vowel* was used as group variable and *F1, F2, F3* and *duration* as dependent variables. From this discriminant function the probability of each production to belong to a certain phoneme was calculated. Table 5.3 presents the results. The table shows mean probabilities over speakers of productions to be classified as a certain phoneme. The lines of the table correspond to the intended phoneme identity, the columns to the classification according to the discriminant function. For example, the 0.913 in column 3, first line, means that the mean probability of the productions of /i/ to be classified as /i/ is 0.913. The 0.031 in the fourth column means that the probability of /i/ productions to be classified as /e/ is 0.031. Bold numbers give correct classifications. The great majority of productions can thus be distinguished by the four parameters: Bold numbers are usually close to 1, the other numbers are close to 0. Some productions, however, are still misclassified, especially productions intended to be /ɪ/. Furthermore, in line with the results from the overlap of the dispersion ellipses, the contrast between /i/ and /e/ is not always sufficient, and there is also some misclassification of /o/ and /u/. Looking at the bold numbers only one can see that for all vowels they are higher in the pre-perturbed session (E1/1) than in the first perturbed session (E1/2) and they drop even more in E1/3. Afterwards, however, they rise. These differences are very small but still consistent, so it seems that at perturbation onset the acoustic differences between the vowels are smaller than in the unperturbed session, but after a few days of adaptation the distinctions return to their original quality. Thus, there is a development towards greater distances in the acoustic space. However, this development exists for the complete complex of acoustic characteristics analysed here and not only for a single formant or a pair of parameters (e.g. F2-F3 space, F2-duration etc.).

Table 5.3: Mean classification probabilities over speakers. Column 1: session in which the sound was produced, column 2: sound which was intended to be produced, columns 3 to 8: classification probabilities. Correct classifications in bold.

		classified as					
session	sound	i	e	y	o	u	ɪ
E1/1	i	**0.913**	0.031	0.000	0.000	0.000	0.056
E1/1	e	0.013	**0.883**	0.000	0.000	0.000	0.104
E1/1	y	0.000	0.000	**0.988**	0.000	0.000	0.011
E1/1	o	0.000	0.000	0.000	**0.980**	0.020	0.000
E1/1	u	0.000	0.000	0.000	0.019	**0.981**	0.000
E1/1	ɪ	0.068	0.075	0.000	0.000	0.000	**0.858**
E1/2	i	**0.849**	0.032	0.000	0.000	0.000	0.119
E1/2	e	0.027	**0.863**	0.001	0.000	0.000	0.109
E1/2	y	0.007	0.000	**0.987**	0.000	0.004	0.001
E1/2	o	0.000	0.000	0.000	**0.953**	0.047	0.000
E1/2	u	0.000	0.000	0.013	0.058	**0.929**	0.000
E1/2	ɪ	0.098	0.086	0.000	0.000	0.000	**0.815**
E1/3	i	**0.778**	0.081	0.000	0.000	0.000	0.141
E1/3	e	0.045	**0.803**	0.000	0.000	0.000	0.152
E1/3	y	0.000	0.000	**0.973**	0.000	0.000	0.027
E1/3	o	0.000	0.000	0.000	**0.954**	0.046	0.000
E1/3	u	0.000	0.000	0.000	0.059	**0.941**	0.000
E1/3	ɪ	0.105	0.129	0.002	0.000	0.000	**0.764**
AU1	i	**0.89**	0.033	0.000	0.000	0.000	0.076
AU1	e	0.031	**0.887**	0.000	0.000	0.000	0.082
AU1	y	0.000	0.000	**0.992**	0.000	0.000	0.008
AU1	o	0.000	0.000	0.000	**0.970**	0.030	0.000
AU1	u	0.000	0.000	0.000	0.027	**0.973**	0.000
AU1	ɪ	0.042	0.072	0.000	0.000	0.000	**0.886**
E2/1	i	**0.910**	0.025	0.000	0.000	0.000	0.064
E2/1	e	0.012	**0.874**	0.000	0.000	0.000	0.114
E2/1	y	0.000	0.000	**0.998**	0.000	0.000	0.002
E2/1	o	0.000	0.000	0.000	**0.959**	0.041	0.000
E2/1	u	0.000	0.000	0.000	0.032	**0.968**	0.000
E2/1	ɪ	0.052	0.102	0.000	0.000	0.000	**0.846**
AU2	i	**0.914**	0.026	0.009	0.000	0.000	0.051
AU2	e	0.035	**0.904**	0.000	0.000	0.000	0.061
AU2	y	0.000	0.000	**0.993**	0.003	0.004	0.000
AU2	o	0.000	0.000	0.000	**0.957**	0.043	0.000
AU2	u	0.000	0.000	0.000	0.030	**0.970**	0.000

AU2	ɪ	0.060	0.063	0.000	0.000	0.000	**0.878**
E3/1	i	**0.922**	0.019	0.000	0.000	0.000	0.059
E3/1	e	0.022	**0.840**	0.000	0.000	0.000	0.138
E3/1	y	0.000	0.000	**1.000**	0.000	0.000	0.000
E3/1	o	0.000	0.000	0.000	**0.978**	0.022	0.000
E3/1	u	0.000	0.000	0.000	0.020	**0.980**	0.000
E3/1	ɪ	0.055	0.100	0.000	0.000	0.000	**0.845**
E3/2	i	**0.909**	0.033	0.000	0.000	0.000	0.058
E3/2	e	0.023	**0.879**	0.000	0.000	0.000	0.098
E3/2	y	0.000	0.000	**1.000**	0.000	0.000	0.000
E3/2	o	0.000	0.000	0.000	**0.978**	0.022	0.000
E3/2	u	0.000	0.000	0.000	0.030	**0.970**	0.000
E3/2	ɪ	0.041	0.104	0.000	0.000	0.000	**0.855**

5.5 Conclusion

The aim of the analysis presented here was two-fold, namely to investigate the classifiability of the vowels and to investigate the development of acoustic parameters (formant values and durations) over time. With respect to the first question it was found that single formants, 2D formant regions and formant-duration regions overlap for the phonemes when all sessions are discussed at the same time. When data are split according to session, however, the great majority of phoneme distinctions becomes clear by just regarding the formant values. Further distinctions are possible when duration is considered (e.g. /o/ vs. /u/ for speaker SK). A discriminant analysis has shown that most of the productions are classified correctly when data are split according to sessions and all four parameters (F1, F2, F3 and duration) are considered. Overlap occurs predominantly for the lax vowel /ɪ/.

Informal perception tests showed that the wrongly classified productions could in fact be correctly classified perceptually, thus further acoustic parameters should be involved, as for example the formant contour, f0 of the syllable structure.

With regard to the development of the values over time a development towards the original formant or duration values cannot be found. However, the discriminant analysis has shown that in the early adaptation phase there is a development towards greater distances between the vowels when all four characteristics are taken into account.

The 2D formant plots suggest that the vowel subregions vary over sessions, however, it is difficult to find a direction of variation. It is generally not the case that, for example, a vowel has certain formant values at perturbation

onset and afterwards there is a development into a certain direction (for an exception see chapter 6). Furthermore, the vowel formants still vary in the later sessions when the distances between the vowels stay constant.

The results could thus be interpreted in the framework of the H & H-Theory (Lindblom [1988], Lindblom [1990]). Invariants (even invariant regions) would thus not exist, but utterances vary for different communicational situations. The listener takes all the available information (linguistic and extralinguistic) in order to understand the utterance, in the present case he would probably take information from the surrounding sounds. Lindblom [1996] discusses how it is possible to recognise stops in different vowel surroundings by using the formant information of surrounding vowels. In the present case it could be the other way round: Information from the surrounding consonants is taken in order to interpret the vowel, or else, information from other vowels is used to set up the framework within which the production is interpreted.

Another possibility, which does not contradict the interpretation in the framework of the H & H- Theory, would be that speakers do not try to reach a certain perceptual target in either the acoustic or the articulatory domain but that they try out different strategies to produce a vowel in order to find an optimal way to produce the sound and that the formant values therefore vary inconsistently.

Coming back to the question whether support for articulatory or acoustic perceptual primitives can be found the fact that the formant values do not develop towards the values measured in the unperturbed session speaks against acoustic speech production tasks, except if one interprets the data in the framework of the H & H-Theory. On the other hand, the fact that F1 and F2 do not vary consistently either leads to the assumption that there is no consistent development in the articulatory domain either.

Chapter 6

Distance in articulatory and acoustic space

Chapter 5 has shown that nearly all vowel productions can be correctly classified according to acoustic parameters even immediately after perturbation onset. In spite of the high classification scores, especially from session AU1 onwards, which suggest a certain stability of the acoustic characteristics, the formant values still vary over time. Obviously, speakers change their articulation even when the acoustic targets are already reached. For the two front vowels /i/ and /e/ a rise of F2 over the adaptation time could be found. It was the only case where a formant varied in any consistent manner. The present chapter aims at finding a possible reason for this rise in F2. Therefore, the articulatory positions and the formant values of the two sounds will be investigated.

After a brief discussion of absolute positional measurements, the articulatory analysis will proceed with measurements of vowel positions relative to each other. There are two reasons for that, a methodological one and a more theoretical one. First, since the sensors were glued to the tongue anew on every recording day one cannot be absolutely sure that the same position on the tongue is measured. If - in absolute terms - a sensor is more advanced in one session than in another one this can well be due to a sensor which is glued to a more anterior part of the tongue. When the distance between positions of the same sensor in two different vowels is measured the influence of the exact positioning of the sensor on the tongue does not matter. The second reason regards the intentions the speakers might have in changing the articulatory position. A possible reason could be to increase the articulatory or acoustic space in between two vowels. Therefore, it seems to be more useful to discuss articulatory and acoustic distances between vowels rather than absolute values.

6.1 Methods

As a first step the tongue dorsum sensor positions at the articulatory target positions (cf. section 4.8.1) of the vowels were plotted for all sessions and the result was investigated visually. For all seven speakers the back vowel positions were very stable over the sessions, but the target positions of the front vowels were lowered when the prosthesis was inserted. The horizontal position of the front vowels changed as well, but this change was not the same for the seven speakers. For the majority of speakers the tongue was retracted immediately after perturbation onset, afterwards speakers forwarded the position of the front vowel gradually, sometimes up to the point where the initial horizontal position was reached. The plots suggested that the articulatory distances among the front vowels stayed the same over the adaptation period, but the distances between front and back vowels changed.

In a second step, distances in articulatory and acoustic space were measured. Distances were investigated for /i/ as compared to /e/, /i/ as compared to /u/ and /e/ as compared to /o/. These pairings were chosen for three reasons. The front vowel-back vowel pairing was selected because the results from the positional analysis similarly to the F2 results from the preceding chapter suggested that there could be an interesting change. The /i/-/e/ pairing was chosen because it is a more critical contrast and could therefore be interesting. For each pairing there is a corresponding acoustic parameter (F1 for the front vowel contrast, F2 for the front-back vowel contrast) for the articulatory dimension investigated. Comparisons were carried out for vowels which were produced closely in time. For example, the first production of /i/ in session E1/1 was compared to the first production of /u/, then the second production of /i/ was compared to the second production of /u/ and so forth.

Articulatory distance was measured as the Euclidean distance between the sensor positions of closely in time produced /i/ and /e/, /i/ and /u/ and /e/ and /o/. Acoustic distance was measured as the difference in F1 between the two front vowels and as the difference in F2 between a *front vowel-back vowel* pair the two productions of which were produced closely in time.

6.2 Results

Since especially the distinctions between the front vowels are critical (cf. chapter 5), one could expect that there will be an increase in distance between these close vowels. In line with the preliminary investigations, however, the changes in acoustic and in articulatory space were minimal and no consistent

development over the sessions could be found. A more detailed presentation of the results will therefore be abstained from. Instead, the presentation of the results will concentrate on the front-back vowel pairs.

Again in line with the results from the preliminary investigation of positional plots, the majority of speakers showed a decrease in articulatory and acoustic space which was followed by an increase in later sessions. Figure 6.1 shows the development of articulatory and acoustic distance for the two vowel pairs over the adaptation time for five speakers. The results for the two further speakers, who do not show the pattern, can be found in tables 6.1 and 6.2. The upper two subfigures in figure 6.1 give the results for the intermediately high vowels, the lower two the ones for the high vowels. The figures on the left show the articulatory distance, the figures on the right the acoustic distance. Each group of bars corresponds to one speaker, each bar gives the mean distance within a session. Bars are in chronological order.[1] Each subfigure will now be discussed in detail.

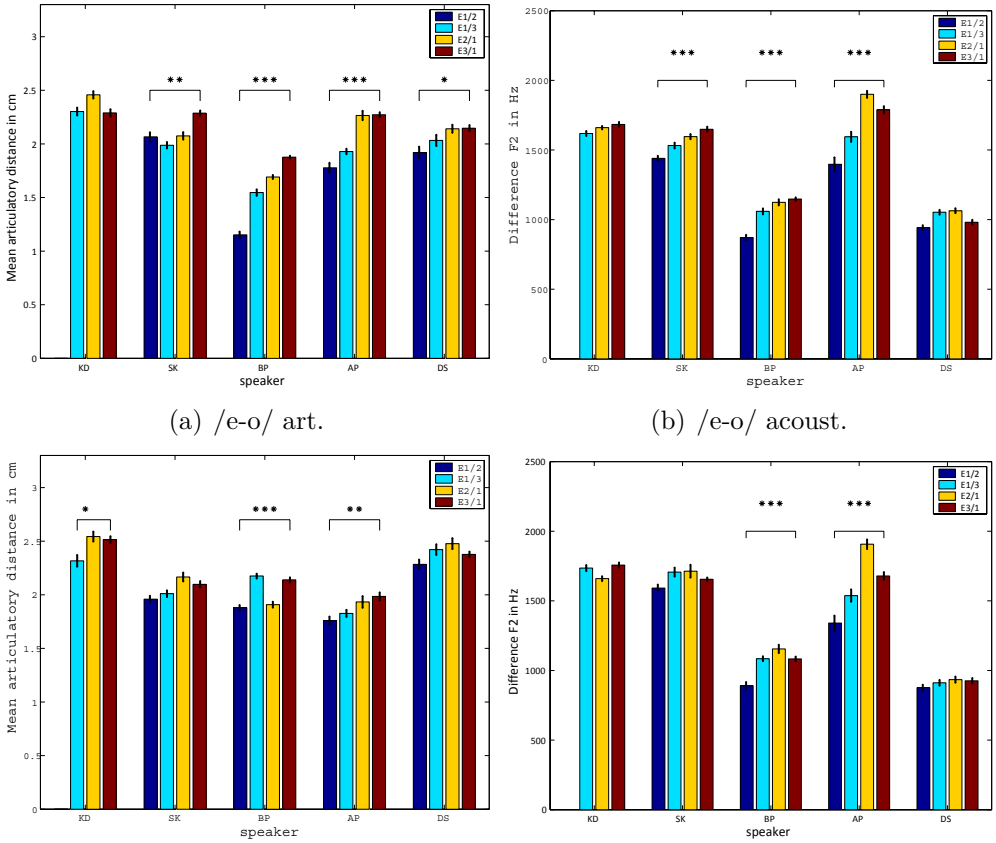
Articulatory distance between /e/ *and* /o/ *(upper left).* For all speakers shown in the figure the articulatory distance is low in the first perturbed session (E1/2, for speaker KD E1/3, first bar). In the later sessions it increases with two exceptions: KD where there is a decrease in session E3/1 after an earlier increase, and a small further decrease in session E1/3 for speaker SK.

Table 6.1 (upper half) gives the means and standard deviations of the articulatory distance for all sessions. The two speakers without a decrease-increase pattern are AP with an increase over all sessions and TP with an initial decrease but an oscillating pattern afterwards. For DS, things are not very clear since the articulatory distance stays about the same in the first perturbed session as in the unperturbed one.

A repeated measures ANOVA for data split by speaker with session as factor and articulatory distance as variable showed that there is a significant influence of the session on the articulatory distance (cf. table A.7, p. 148). In order to investigate whether the articulatory distance between session E1/2 and E3/1 increases significantly, Tamhane T2 post-hoc tests were calculated. Except for speaker KD the difference is significant for all speakers shown in figure 6.1 (see significance levels given in the figure).

Articulatory distance between /i/ *and* /u/ *(lower left subfigure).* The figure shows the articulatory distance for the high vowel pair. In general, there is the same tendency as for the intermediately high vowels, an increase over the adaptation time, but this increase is less clear since it is not monotonic. For speaker BP there is an increase in distance in session E1/3 (second bar), then there is a drop in session E2/1 (third bar), followed by a rise in session

[1] The pre-perturbed session is not shown in order not to overload the figure.



(a) /e-o/ art. (b) /e-o/ acoust.

(c) /i-u/ art. (d) /i-u/ acoust.

Figure 6.1: Development of the articulatory (left) and acoustic (right) distance between /e/ and /o/ (upper half) and /i/ and /u/ (lower half) over the adaptation time for the five speakers for whom there is a decrease of distance at perturbation onset. Each bar shows the mean for one session. First bar of each group of bars: E1/2, second bar of a group: E1/3, third bar of a group: E2/1, fourth bar of a group: E3/1. Note that session 1 (first bar) is missing for speaker KD. Error bars show standard error. Horizontal brackets above the bars denote cases with significant increases from the first to the last perturbed session. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

E3/1 (brown bar). For speaker SK the distance is a little smaller in session E3/1 than in session E2/1, the same is true for speakers KD and DS. A significant difference between the initial and the final perturbed session can only be found for three speakers (see significance levels in the figure). A repeated measures ANOVA has shown that the influence of the session on

Table 6.1: Means and standard deviation (in parentheses) for articulatory distance between /e/ and /o/ (upper half) and /i/ and /u/ (lower half). Session E1/2 is missing for two speakers (cf. section 4.4).

speaker	E1/1	E1/2	E1/3	E2/1	E3/1
TP	2.97 (0.11)	2.48 (0.13)	2.89 (0.23)	2.67 (0.17)	2.02 (0.19)
KD	2.76 (0.12)		2.30 (0.17)	2.46 (0.15)	2.29 (0.15)
OP	3.03 (0.28)		3.17 (0.19)	3.03 (0.16)	2.63 (0.16)
SK	2.44 (0.08)	2.07 (0.19)	1.99 (0.14)	2.08 (0.16)	2.29 (0.11)
BP	1.68 (0.12)	1.15 (0.14)	1.55 (0.14)	1.70 (0.09)	1.88 (0.06)
AP	2.59 (0.12)	1.78 (0.20)	1.93 (0.11)	2.27 (0.19)	2.27 (0.10)
DS	1.91 (0.14)	1.92 (0.24)	2.03 (0.23)	2.14 (0.17)	2.15 (0.13)
TP	2.96 (0.14)	2.61 (0.19)	2.86 (0.19)	2.88 (0.18)	1.91 (0.17)
KD	2.84 (0.11)		2.32 (0.24)	2.54 (0.21)	2.52 (0.14)
OP	2.72 (0.18)		2.92 (0.19)	2.91 (0.18)	2.48 (0.11)
SK	2.46 (0.10)	1.96 (0.14)	2.01 (0.14)	2.17 (0.18)	2.10 (0.13)
BP	2.45 (0.10)	1.88 (0.11)	2.18 (0.10)	1.91 (0.13)	2.14 (0.10)
AP	2.50 (0.18)	1.76 (0.17)	1.83 (0.15)	1.93 (0.25)	1.99 (0.17)
DS	2.44 (0.14)	2.28 (0.20)	2.42 (0.22)	2.48 (0.23)	2.38 (0.12)

the articulatory distance is in all cases significant (cf. table A.7).

Table 6.1, lower half, gives means and standard deviations for all speakers. Again, the decrease-increase pattern cannot be found for the two speakers not shown in the figure, TP and OP. For speaker TP there is an increase in distance until session E2/1, but afterwards there is a drop. For speaker OP there is also an increase followed by a decrease. However, the pattern is clear for speaker DS now since there is a drop in session E1/2 as compared to E1/1.

Acoustic distance between /e/ *and* /o/ *(upper right subfigure).* The figure shows the difference in F2 between the two intermediately high vowels. Comparing the mean values over sessions there is an increase in distance similarly to the one for the articulatory data. A repeated measures ANOVA for data split by speakers has shown that there is an influence of the session on the acoustic distance (cf. table A.7). Tamhane T2 post-hoc tests show that the difference between session E1/2 and E3/1 is significant for three of the five speakers.

Table 6.2, upper half, gives the means and standard deviations for each session and each speaker. Similarly as for the articulatory data, the decrease-increase pattern cannot be found for speakers OP and TP.

Acoustic distance between /i/ *and* /u/ *(lower right subfigure).* Similarly

Table 6.2: Means and standard deviation (in parentheses) for acoustic distance between /e/ and /o/ (upper half) and /i/ and /u/ (lower half).

speaker	E1/1	E1/2	E1/3	E2/1	E3/1
TP	1293 (49)	1296 (57)	1256 (54)	1313 (37)	1230 (61)
KD	1660 (62)		1619 (79)	1660 (55)	1684 (79)
OP	1981 (76)		1923 (161)	1863 (109)	1815 (76)
SK	1600 (53)	1440 (78)	1532 (93)	1596 (84)	1649 (80)
BP	1146 (87)	871 (83)	1059 (96)	1124 (95)	1148 (50)
AP	1641 (213)	1397 (224)	1595 (164)	1900 (112)	1788 (116)
DS	1039 (88)	942 (76)	1053 (75)	1063 (78)	981 (75)
TP	1282 (69)	1091 (110)	1175 (100)	1312 (78)	1061 (128)
KD	1692 (90)		1735 (97)	1660 (76)	1757 (85)
OP	2229 (78)		2114 (76)	2043 (84)	2024 (90)
SK	1834 (151)	1592 (108)	1706 (146)	1713 (206)	1655 (59)
BP	1182 (71)	890 (121)	1084 (79)	1155 (134)	1083 (77)
AP	1476 (164)	1340 (241)	1538 (196)	1907 (154)	1679 (128)
DS	998 (93)	876 (96)	912 (94)	934 (97)	926 (83)

as for the articulatory distance, an increase can be seen but it is not monotonic. Even if the influence of the session on the acoustic distance is again significant, the initial and the final perturbed session differ significantly for only two speakers. Means and standard deviations can be found in table 6.2, the results of the ANOVA are in table A.7.

6.3 Conclusion

Whereas there are no changes in articulatory and acoustic distance between the two front vowels, changes could be found for the distinction between front and back vowels. The majority of speakers produces a smaller articulatory and acoustic distance at perturbation onset as compared to the pre-perturbed session, and this distance increases over the adaptation time. The increase is equally consistent in the articulatory as in the acoustic domain, but clearer for the intermediately high vowels than for the high vowels.

Since it is not the case that the pattern is clearer either in the acoustic or in the articulatory domain the interpretation of this result is not straightforward. If speakers adapt towards an acoustic target they might try to reach the original F2 value. If speakers adapt towards articulatory targets they might move towards the original place of articulation.

A question which arises is why the development towards greater distances

exists for these already very clear contrasts between front and back vowels but not for the critical contrast between /i/ and /e/. One explanation could be that speakers easily manage to increase the first distance but it is just too difficult to increase the distance between the two front vowels. Another explanation could lie in the immediate compensatory behaviour. Since there are no central vowels in German (at least not in stressed position) speakers can easily move their tongue "out of the way" from the prosthesis by retracting it, without running the risk to produce a sound which is not correctly perceived. So they might be more prepared to accept an articulatory change at perturbation onset because the acoustic output is still not in the region of another phoneme. Thus, there is just a greater deviation at perturbation onset which is corrected afterwards.

A further question is why the development towards a greater distance is clearer for the intermediately high vowels than for the high vowels. This could be due to general difficulties during the production of /i/ (as compared to /e/) which does not allow for very much articulatory variability without changes in the acoustics. Speakers, even if they aim at producing a greater distance, might not be successful in producing it correctly.

The reason for the initial retraction in /i/ and /e/ could be an acoustic one. Speakers might try to prevent having linguo-palatal contact in order to preserve the basic vocalic characteristics. When the constriction is too small this should result in frication, which is not typical for German vowels.

An articulatory reason could be that speakers estimate the constriction degree by linguo-palatal contact. When the artificial palate is inserted they should, without retraction and lowering, have more linguo-palatal contact. Since the palate is posteriorily higher, retracting the tongue should result in a contact pattern which is similar to the initial one while the same degree of opening of the jaw can be kept. Thus, initially, especially when no auditory feedback is available, speakers could make use of an articulatory representation in form of contact patterns in order to produce the necessary articulatory configuration. Later they might use auditory feedback in order to produce the original formant pattern.

Thus, a hypothesis, which will be explored more and more in the due course of this study, is that initially speakers are using articulatory representations of the sounds (possibly contact patterns), and only afterwards they adapt towards acoustic targets. A second possibility is that the retraction is more or less mechanical and afterwards speakers adapt towards an articulatory target, namely the original constriction location.

Chapter 7

Acoustic characteristics of fricatives

In some of the earlier perturbation studies described in chapter 3 it was found that fricatives are much more difficult to adapt than vowels. In fact, returning to the classification between compensation by reparametrisation, a strategy which uses the learned articulatory strategy and just adapts a few parameters, and compensation by reorganisation, which creates a new strategy (cf. chapter 3), one could even assume that in vowel adaptation a reparametrisation might be sufficient. For consonant adaptation, however, previous results show that a new articulatory strategy has to be developed. This might take several days or weeks. Similarly as for the vowels, however, speakers' efforts should be driven towards the creation of separate regions in acoustic space. Moreover, it is possible that they try to produce fricatives which are acoustically similar to the ones produced in the unperturbed session.

As described in section 4.7.3 a number of parameters describing the fricative spectrum (cog, dispersion, skewness, kurtosis, a-slope and b-slope) were calculated. An inspection of each single parameter in general did not show consistent developments, at least not across speakers. The only exception is a decrease of the COG in the first perturbed sessions which is consistent over the majority of speakers. This exception will be discussed separately in chapter 9. The present chapter is concerned with the more global development of the acoustic characteristics over the complete adaptation time. In order to include the information of all the parameters in the analysis, two discrimininant analyses were carried out, the first in order to distinguish between phonemes (cf. section 7.1), the second in order to distinguish between unperturbed and perturbed productions (cf. section 7.2). The audio sessions were excluded from the analysis because the EMA setup had an influence on

the fricative acoustics so that audio and EMA sessions could not be grouped together. In the second analysis this exclusion is absolutely necessary since the discriminant function, which is calculated on EMA-data, will not be useful for a classification of data from the audio-only recordings. However, even in the first analysis, where a separate function is calculated for each session, the two audio sessions were excluded because the finding would probably be that the classification is better in the pure audio sessions than in the EMA-sessions. The result would thus not contribute to the question whether the classification becomes clearer over the adaptation time. The analysis here is restricted to acoustics. For information on articulatory data for the fricatives the reader is referred to chapter 9.

7.1 Acoustic distinction between phonemes

The analysis of the vowels has already shown that the regions in acoustic space vary over sessions, and that this leads to overlap of the acoustic regions of phonemes across sessions. Looking at single parameters (cog, dispersion etc.) the same could be found for the fricatives. Data were therefore split by session and speaker in order to investigate whether phonemes can be reliably distinguished from each other at least within a session. Then a discriminant analysis was calculated with the six spectral parameters. The groups in each discriminant function were the different sounds (/z, s/, initial /ʃ/, /ç/ and /x/). The probability of each realisation to belong to the group of each sound was calculated (cf. section 4.7.4).

For the further analysis, means of the probabilities of the productions to belong to a phoneme were calculated over all speakers. Table 7.1 gives the results. The table has the same structure as the one for the vowels in chapter 5. The first column gives the session, the second the sound which was intended to be produced. Columns 3 to 7 give the mean probability of these productions to belong to a certain phoneme. For example, the mean probability of all /z/-productions to be (correctly) classified as /z/ is 0.855. The probability to be (wrongly) classified as /s/ is 0.145. The correct classifications are marked in bold. Most productions can thus be correctly classified by the discriminant function. Many of the incorrectly classified productions of alveolar fricatives are classified as the concurrent fricative with the opposite voicing characteristic. This is probably due to the high pass filtering carried out before the calculation of the spectral parameters which eliminated the voicing characteristics in /z/. The articulatory strategy chosen by the subjects thus permitted to maintain the distinction between consonants articulated in different regions of the vocal tract.

Table 7.1: Mean classification probabilities over speakers. Column 1: session in which the sound was produced, column 2: sound which was intended to be produced, columns 3 to 7: classification probabilities. Correct classifications in bold.

		classified as				
session	sound	z	s	∫	ç	x
E1/1	z	**0.855**	0.145	0.000	0.000	0.000
E1/1	s	0.078	**0.922**	0.000	0.000	0.000
E1/1	∫	0.000	0.000	**0.990**	0.010	0.000
E1/1	ç	0.000	0.000	0.001	**0.999**	0.000
E1/1	x	0.000	0.000	0.000	0.000	**1.000**
E1/2	z	**0.855**	0.145	0.000	0.000	0.000
E1/2	s	0.147	**0.853**	0.000	0.000	0.000
E1/2	∫	0.000	0.000	**0.984**	0.015	0.001
E1/2	ç	0.000	0.000	0.000	**1.000**	0.000
E1/2	x	0.000	0.000	0.022	0.000	**0.978**
E1/3	z	**0.879**	0.121	0.000	0.000	0.000
E1/3	s	0.104	**0.896**	0.000	0.000	0.000
E1/3	∫	0.004	0.000	**0.983**	0.008	0.005
E1/3	ç	0.000	0.007	0.005	**0.988**	0.000
E1/3	x	0.000	0.000	0.000	0.000	**1.000**
E2/1	z	**0.881**	0.119	0.000	0.000	0.000
E2/1	s	0.086	**0.914**	0.000	0.000	0.000
E2/1	∫	0.000	0.000	**1.000**	0.000	0.000
E2/1	ç	0.000	0.000	0.000	**1.000**	0.000
E2/1	x	0.000	0.000	0.000	0.000	**1.000**
E3/1	z	**0.875**	0.125	0.000	0.000	0.000
E3/1	s	0.100	**0.900**	0.000	0.000	0.000
E3/1	∫	0.000	0.000	**1.000**	0.000	0.000
E3/1	ç	0.000	0.000	0.003	**0.997**	0.000
E3/1	x	0.000	0.000	0.000	0.000	**1.000**
E3/2	z	**0.863**	0.137	0.000	0.000	0.000
E3/2	s	0.126	**0.874**	0.000	0.000	0.000
E3/2	∫	0.000	0.000	**0.992**	0.008	0.001
E3/2	ç	0.000	0.000	0.006	**0.994**	0.000
E3/2	x	0.000	0.000	0.000	0.000	**1.000**

Looking at the development of the identification scores by comparing different sessions, contrary to the vowels, it is not easy to find an improvement

over the perturbed sessions. There might be one in /s/ from session E1/2 (0.853) to E2/1 (0.914), but then the score drops again to 0.900 in session E3/1. For /x/ there is a drop in the first perturbed session, afterwards the sound is always correctly classified.

The results thus show that nearly all the productions can be correctly classified by acoustic characteristics already very early and even in the session without auditory feedback. If speakers aim is to produce fricatives which can be correctly classified they reached it already very early.

7.2 Development toward unperturbed speech

This section tries to answer the question whether, apart from producing fricatives which can unambiguously be assigned to a phoneme, speakers over time make their productions more similar to the unperturbed ones. For the vowels it has been found that this is not the case.

In order to investigate this question, data were split according to sound and speaker. The first unperturbed and the first perturbed session were defined as groups and, via a discriminant function, the probability of the productions of all the sounds of all sessions to belong to one of these two groups was calculated. If there is a development towards the original acoustic characteristics the probability to belong to the unperturbed session should at first be very low and increase over the adaptation time.

Figure 7.1 gives the mean probability over speakers of the productions of each session (abscissa) to belong to the group of unperturbed sounds (ordinate). Each line represents one sound. For example, looking at the dark grey dash-dotted line representing /z/ (dark grey, dash-dotted) the probability of the productions of the first subsession (the unperturbed productions) to belong to the class of unperturbed productions is about 1. This is not surprising since all these productions are unperturbed. Equally unsurprising, the value of the second session is nearly 0 since all these productions are perturbed and created the second group in the discriminant analysis.[1]

Afterwards, from session E1/3 onwards, however, the values rise. This means that, judging from the measured parameters the perturbed productions become more like the unperturbed productions. The last bar gives the values of the final unperturbed session. The probability of these productions to belong to the unperturbed group is below 1. This means that the

[1]The mean values are not exactly 1 and exactly 0 since the productions from sessions E1/1 and E1/2 were classified as well and there were a number of untypical ones which did not have a probability of 1 or 0, respectively.

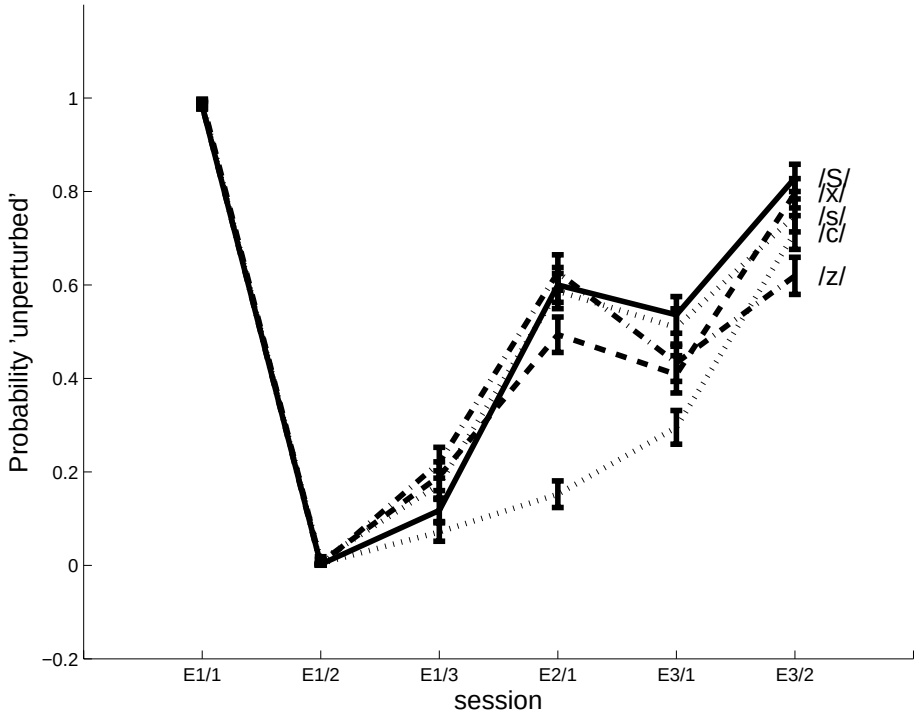


Figure 7.1: Mean probability over speakers to belong to the group of unperturbed productions. Each line gives the results for one sound. Error bars show standard error. Grey shades and line styles as indicated on the right of the plots: light grey, dotted: /ç/, black, dotted: /s/, light grey, solid: /ʃ/, black, dashed: /x/, dark grey, dash-dotted: /z/.

adaptation has changed the characteristics of the unperturbed articulation somehow.

There are differences with respect to the sound. The palatal fricative (light grey, dotted) improves steadily but slowly. The voiced alveolar fricative (dark grey, dash-dotted) appears to be the most difficult one: It does not reach as high probability values as the other sounds. A repeated measures ANOVA for data split by speaker and item shows a significant influence of the session on the probability in all cases. In order to judge whether there was a significant improvement of the productions, Tamhane T2-posthoc tests were calculated. A *significant improvement* was defined as a significant difference somewhere from session E1/3 to session E3/1 with the later session having a higher probability value than the preceding session. In principle, one could have regarded the interval from session E1/2 to session E3/1, but since the probability function was calculated from the data in session E1/2, it was

decided to be more conservative and choose an "independent" session. For the speakers for whom there was no session E1/2 and for whom consequently the probability function had been calculated from session E1/3, a significant increase was searched for in the interval between E2/1 and E3/1.

For the speakers with a session E1/2 there was a significant increase in all cases except for /x/ produced by speakers TP, SK and AP, for /ʃ/ produced by speakers TP and SK, for /ç/ produced by speakers BP, DS and SK, for /s/ produced by speaker DS and for /z/ produced by speaker TP. For the two speakers for whom there was no session E1/2 a significant increase could nearly never been found, but this is probably because it would have had to occur in this rather short interval between sessions E2/1 and E3/1.

7.3 Conclusion

The first part of the analysis presented here was concerned with the classifiability of the fricative productions under perturbation. The main result was that, within a session and for a single speaker, nearly all the productions can be correctly classified by acoustic parameters immediately after perturbation onset. This result is comparable to the ones found for the vowels. A difference between the vowels and the fricatives is that for the vowels the classification became worse in session E1/3 as compared to session E1/2 whereas the fricative productions stayed at about the same level.

In contrast to the vowels, where an inconsistent development of the acoustic parameters over the adaptation time was found after the acoustic target regions had been reached, a systematic development could be found for the fricatives. Speakers produce fricatives which are acoustically more and more like the ones of the unperturbed session.

The results show that, comparable to the results presented for example in Hamlet and Stone [1978] and Baum and McFarland [2000], fricative adaptation takes rather long.

With respect to the speech production tasks the results suggest that even very subtle acoustic characteristics could belong to the phoneme representations.

It is hard to allocate the improvement to a certain acoustic parameter. Furthermore, in contrast to the vowels, where speakers showed about the same behaviour, speakers differ a lot in their strategies, but also in their capabilities to adapt the fricatives.

The present chapter has dealt with the acoustic characteristics of fricatives only, and, as indicated, the articulation will be discussed in chapter 9. The main topic of this later chapter, however, is the influence of auditory

feedback and it will be preceded by a discussion of the influence of auditory feedback in vowel production (chapter 8).

Chapter 8

The influence of auditory feedback in vowel production

A lot of evidence has been found that auditory feedback is important for speech acquisition and maintenance. Speakers who lose hearing are able to produce normal speech for years (Perkell et al. [2000]). Afterwards, however, speech deteriorates due to anatomical changes in the vocal tract which speakers do not compensate for (Waldstein [1990]).

This capability to use speech even after hearing loss has often been explained with the existence of an internal model (e.g. Guenther [1995], Perkell et al. [2000], Jordan [1996]). As described in chapter 2 already, an internal model is a mapping of motor components (e.g. motor commands) to an acoustic output that exists in the speaker's brain. Following this assumption, online auditory feedback is not absolutely necessary for speech production because speakers, with the help of their internal model, know what to do in order to produce a certain acoustic output[1]. However, when the vocal tract changes, the internal model does not correctly predict the acoustic output anymore and auditory feedback is needed in order to readjust it.

Many experiments supporting acoustic phoneme representations dealt with the adaption to a manipulation of the auditory feedback in order to find out whether auditory feedback is used online in order to adapt and/or whether it is used to maintain the internal model. Houde and Jordan [2002], for example, shifted the feedback of whispered productions of /ɛ/ either towards /a/ or towards /i/. In dependence on the direction of the shift speakers compensated differently by a change in articulation, and these changes were retained when speakers whispered with auditory feedback blocked by white

[1] assuming acoustic representations as it is done in the internal model approaches cited above

noise. This underlines the use of auditory feedback for the maintenance of the internal model.

Other experiments show that speakers do not only adapt formant frequencies, but also f0. Jones and Munhall [2000] manipulated f0 before feeding it back in real time. For one group of subjects f0 was shifted upwards, for another group it was shifted downwards. Speakers compensated in the opposite direction and showed an after effect. Same as in the experiment by Houde and Jordan, the after effect shows that the production is not only dependent on auditory feedback but also on an articulatory representation, which, however, is learned from acoustics and disappears once it is not appropriate any more for the acoustic objectives.

In the experiment reported in Purcell and Munhall [2006a] spontaneous adaptation was tested. Subjects produced sustained /ɛ/. At 300 ms F1-feedback was shifted upwards towards /æ/ or downwards towards /i/. In order to prevent speakers from adapting (in the sense of a modification of the internal mapping between motor commands and acoustics) most trials were carried out without feedback manipulation. It was found that the adaptation was rather small as compared to the change in F1 in the auditory feedback. Purcell and Munhall suggest that this could be due to the bone conduction which influences the auditory feedback.

In a further experiment, Purcell and Munhall [2006b] investigated the compensation threshold. Speakers produced the word "head" and the F1-feedback of the vowel was gradually manipulated. Speakers did not compensate when the change was below 60 Hz, and when they adapted the compensation was only partial. There was an after effect which was, however, independent of the duration speakers had been exposed to the shifted F1 frequencies, and the after effect did not occur in a direction that supports the hypothesis of a redefinition of an internal model. The results are interpreted as support for the set up of new articulatory targets which, however, slowly disappear due to a memory decay effect.

Jones and Munhall [2003] is the only previous experiment where compensation to a structural modification of the vocal tract was investigated under auditory feedback masking conditions. As described in chapter 1 already, in this experiment the articulation of /s/ was perturbed by extending the length of the maxillary incisors. Speakers' auditory feedback was alternately masked and non-masked. Speakers started compensation only when auditory feedback was available. For the correct production of /s/ under perturbation auditory feedback therefore seems to be an absolute condition.

The results from these previous studies suggest for the experiment discussed here that speakers might improve their productions in the session when auditory feedback is available as compared to the preceding session

when no auditory feedback is available. Even when speakers manage to reach the acoustic target regions in session E1/2 they might change their productions to make them more similar to the unperturbed productions when auditory feedback is available. The fact that the distances between the vocalic phonemes become smaller in session E1/3 as compared to session E1/2 (cf. chapter 5) furthermore suggests that speakers at least behave differently when auditory feedback is there.

In the following sections compensation of vowel acoustics in sessions E1/2 and E1/3 will be investigated. At the onset of session E1/2 speakers had no experience in speaking with the palatal prosthesis. However, they could adapt while making use of tactile feedback. Due to the presence of the artificial palate this tactile feedback was of course reduced. The receptors at the palate could not be used in order to track linguo-palatal contact, but the receptors on the tongue provided information.

8.1 Methods and Results

An informal perception test comparing the productions of session E1/2 and E1/3 suggested that the acoustic differences in vowel production between the session with feedback masking and the one without cannot easily be perceived.

The calculations presented in the following sections were carried out for only those five speakers for whom there was a session with auditory feedback masking. At first, general tendencies for results pooled over speakers will be discussed even if these tendencies are in general not statistically significant. Afterwards, data will be split by speaker and statistics will be repeated.

F1 was measured as described in section 4.7.2. Afterwards a number of calculations were carried out with the aim of investigating whether there is an improvement in the session when auditory feedback is available as compared to the session when no auditory feedback is available. Therefore, mean F1-values were calculated for each vowel and each speaker from the around 20 repetitions of the vowel in the pre-perturbed session (E1/1). These values can be regarded as references for the sounds, the acoustic targets the speakers want to reach.

Afterwards, the absolute difference between each perturbed production from sessions E1/2 (with feedback masked) E1/3 (with feedback available) and this reference value was calculated. This value can be interpreted as a measurement of "quality lag" of the respective production. If it is low, the production is very similar to the unperturbed productions, if it is high, it is very different from the unperturbed productions. The reason for taking the

absolute values is that the direction of the deviations was not consistent over speakers and sessions. Figure 8.1 shows the results for each sound pooled over speakers.

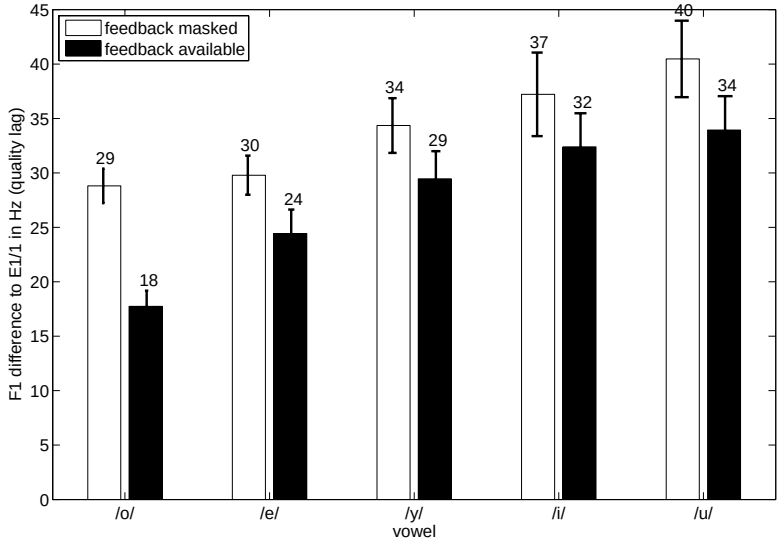


Figure 8.1: "Quality lag" (difference F1 in Hz between the pre-perturbed productions and the perturbed productions) for the session with auditory feedback masked (white) and with auditory feedback available (black) for the five vowels investigated. Error bars show standard error.

In general, the "quality lag" is rather small, the means vary between 18 and 40 Hz, so, in line with the informal perception tests, for the correct classification of the sound the "quality lag" should not play a role. For all items the "quality lag" is higher for the session without feedback (E1/2) than for the session with feedback (E1/3). An ANOVA pooled over speakers showed that this difference between E1/2 and E1/3 is only significant for /o/ ($F(1, 197)=26.325$, $p < 0.001$). For /e/ the result is a little worse ($F(1, 197)=3.490$, $p = 0.063$) Looking at mean values there is more "quality lag" in the high vowels (/y, i, u/) than in intermediately high vowels (/e, o/). High vowels thus seem to be more difficult to produce under perturbation than intermediately high vowels. This could be due to the fact that lower vowels are acoustically more stable than high vowels given the same degree of articulatory variation (Gay et al. [1992]). For example, the tongue height in /a/ can vary a number of millimeters without much change in the acoustics. During the production of /e/ the tongue height can vary less than during /a/, but still more than for /i/ which, for little articulatory variation, soon

becomes either fricativised or /e/-like. Thus, for intermediately high vowels the articulatory-to-acoustic relation is less sensitive than for high vowels and they might therefore be easier to adapt.

A further point will be mentioned in order to show a general tendency which can be seen in figure 8.1. If one takes the difference between the mean values for the two sessions shown in figure 8.1 one gets a measurement which could be interpreted as the "gain", the degree to which a sound "profits" from auditory feedback. This "gain" is always very small. It is highest for /o/ (11), lower for /e/ and /u/ (6), and lowest /i/ and /y/ (5). For the statistical analysis the mean differences between the quality lags for each speaker were calculated and an ANOVA with "gain" as variable and tongue position (high vowels versus intermediately high vowels) as factor was calculated. There is no significant influence of the vowel classification (high - intermediately high) on the "gain" ($F(1, 25)=0.346$, $p = 0.562$). Still, there could be a certain tendency which could be a starting point for further investigations. The intermediately high vowels seem to profit more from auditory feedback than the high vowels; and the back vowels profit more than the front vowels. The reason for this pattern could be palatal contact: There is more linguo-palatal contact in front vowels than in back vowels, and more in high vowels than in low vowels. When there is a lot of contact during a sound it can be produced correctly even when no auditory feedback is available, simply because the speakers can rely on the tongue palate contact pattern associated with this sound. Even if the palate shape is different under perturbation, e.g. because the palate is flatter, the speaker might for example remember that for /i/ bilateral contact is necessary. The speaker could thus try to create this contact pattern, even if this would involve a change in tongue shape (less bunching). On the other hand, if there is less contact, it should be difficult to find the correct tongue height which produces the correct sound because the tongue height cannot be determined by an associated contact pattern. Since the differences in the data are small, however, this can only be seen as a hypothesis.

An explanation for the little improvement in high vowels when auditory feedback becomes available, and at the same time also for the high "quality lag" in these vowels, which goes in the opposite direction, could be that these vowels are more difficult to adapt because, for a changed palate shape, a new tongue shape is necessary in order to produce the correct constriction. For low vowels the tongue shape can stay the same and the sound can still be clearly identified as the required vowel.

In general, the results presented in figure 8.1 can only be seen as a rough tendency since there is a significant influence of the speaker on the "quality lag" so that the data should in fact not be pooled over speakers.

When data were split by speaker, the results became a little inconsistent. In four cases the productions actually became worse when auditory feedback was available: /i, e/ by speaker TP, /i/ by speaker SK, /u/ by speaker DS. An ANOVA with split by speaker and item showed that most differences between the session with feedback masking and the one without feedback masking are not significant. Table 8.1 gives the cases where the differences are significant. Letters in italics mean that the distinction which is significant is not an improvement when auditory feedback is available but that the productions actually become worse.

Table 8.1: Cases in which there was a significant difference in "quality lag" between the session when auditory feedback was masked and the session when auditory feedback was available. Italics: no improvement, but productions become worse.

speaker	vowel	p
TP	/o/	0.000
SK	/i/	*0.009*
BP	/e/	0.000
BP	/o/	0.000
BP	/u/	0.001
DS	/e/	0.006
DS	/y/	0.023
DS	/o/	0.000
DS	/u/	*0.000*

An important point to make is that the improvement from session E1/2 to session E1/3, if it exists, is in most cases not significant. Assuming that a lot of adaptation is necessary for a perturbation as the one investigated here, one can judge that the greatest part of the adaptation is carried out in the session without auditory feedback.

8.2 Conclusion

In summary, the following two tendencies could be found:

- (1) Speakers adapt very well already in the session without auditory feedback. The quality normally improves when auditory feedback becomes available, but the "quality lag" is rather small.

- (2) There seems to be a tendency, although it is not significant, towards a greater "gain" of the auditory feedback for vowels with less linguo-palatal contact than for vowels with a lot of contact. In other words, the productions improve more due to the availability of auditory feedback when there is less contact.

The results are in contrast to the ones found by Jones and Munhall [2003] for their structural perturbation experiment. Speakers in the present experiment already adapt with the use of tactile feedback when no auditory feedback is available. A possible reason could be that Jones & Munhall investigate /s/ whereas the analysis here restricts itself to vowels.

The generally small "quality lag" (below the compensation threshold found by Purcell and Munhall [2006b]) suggests that tactile feedback seems to be sufficient in order to compensate surprisingly well. Speakers thus must dispose of an articulatory phoneme representation, even if it might not be the one which is transmitted to the speaker. This articulatory representation could be a linguo-palatal contact pattern which the speakers try to reproduce when no auditory feedback is available while using tactile feedback. The representation could also be a constriction or a gesture, but, since the tongue and jaw position changes for the vowels, in this case one would have to assume that speakers are able to reparametrise this representation.

Another possibility could be that the feedback masking was not complete and speakers possibly heard themselves during the session with auditory feedback masking via bone conduction. Pörschmann [2000] found that bone conducted sound can have a significant influence on the signal reaching the cochlea. Purcell et al. [2003] found that bone conduction is relatively individual. This could explain why the improvements between the first and the second perturbed session are higher for some speakers than for others.

The second finding, the possibly higher "gain" for intermediately high vowels, can be explained as follows. When there is more linguo-palatal contact the speakers can make use of tactile feedback, which is available in the session without auditory feedback already. The productions are thus already rather good in session E1/2 and improve only little in session E1/3.

Chapter 9

Early adaptation in fricatives

In chapter 7 the development of fricative productions over the adaptation time was discussed. In order to investigate a possible improvement of the productions over time, a discrimininant function was calculated from the pre-perturbed and the first perturbed sessions. Afterwards, all the productions were classified as either unperturbed or perturbed on the basis of these two reference classes. As shown in figure 7.1, p.87, the productions take rather long in order to improve. In session E1/3 maximally 25% of the productions of a sound were classified as unperturbed. The great majority of productions was thus still more similar to the perturbed productions. The analysis presented in chapter 7 did not allow to assess more precisely the differences between sessions E1/2 and E1/3, and consequently the influence of auditory feedback, since one of the sessions was used in order to calculate the discriminant function and the other one was not. There is thus no "background" against which the two sessions could be compared. The analysis presented here will therefore turn to this so far missing point and discuss early adaptation with and without auditory feedback. A further aim of this chapter is to discuss the articulatory correlates of the acoustic changes which were completely missing in chapter 7.

Previous experiments with artificial palates (e.g. Baum and McFarland [1997], Baum and McFarland [2000], Aasland et al. [2006]) showed that the centre of gravity of alveolar fricatives is lower at the beginning of the perturbation but rises with practice. Articulatorily, this lowering of the COG can be explained by a retraction of the constriction and a consecutive lengthening of the front cavity (distance from the constriction to the teeth). This initial retraction which takes place instead of a simple lowering of the tongue or the jaw could have several reasons of which three will be discussed here.

Mechanical move. First, the retraction could be a rather mechanical move "out of the way" because there is something which prevents the speaker from

placing the tongue in the habitual way. A purely mechanical retraction, if it does not correspond to the phoneme representation, should disappear rather soon, maybe even during the first perturbed session (assuming articulatory representations). It should definitely disappear when auditory feedback becomes available because then speakers should notice the lower COG values and correct them.

Linguo-palatal contact pattern. Second, it could be that speakers have an articulatory representation in terms of a linguo-palatal contact pattern and try to realise this pattern under perturbation. At least for the alveolar prosthesis a possible way to reach this contact pattern would be a retraction of the tongue. If speakers retract the tongue in order to reach a certain contact pattern they should be able to find it or at least move somewhere close to it in the first perturbed session and keep it in the session with auditory feedback available. Furthermore, the pattern should be clearer for speakers with an alveolar prosthesis than for the other speakers. A third indicator for the use of a linguo-palatal contact pattern could be that the results differ for sounds which are directly perturbed because their constriction is somewhere at the hard palate, as /s/ and /ʃ/, and sounds which are not directly perturbed, namely /ç/ and /x/.

Acoustic characteristics. Third, it is possible that speakers are unable to produce important characteristics of the fricatives with the tongue in the habitual position and therefore retract it. In order to produce /s/, for example, the tongue blade needs to form a constriction in the alveolar region. The air jet which passes this constriction is afterwards directed against the upper or lower teeth which form a second obstacle in the air passage (Shadle [1985]). This second obstacle gives the sibilant the characteristic high amplitude noise (Shadle [1991]). When the prosthesis is inserted the upper incisors no longer serve as a second obstacle since the prosthesis thickens the alveolar ridge so that it is nearly as thick as the upper incisors are long. The speaker could compensate by using the lower teeth as the only obstacle. This necessitates a high jaw position. If the constriction is formed at the original place, however, and the jaw is high, there should not be enough space left between tongue and palate, since the prosthesis lowers the palate. The only possibility to keep a high jaw position is therefore to retract the tongue towards a region with more space in the vertical dimension. If speakers compensate in this way they should retract the tongue more in session E1/3 when auditory feedback becomes available than in E1/2 because only then they can judge the influence of the jaw position on the acoustic output. Furthermore, if speakers adapt in this way, the jaw should be in a higher position in session E1/3 than in E1/2.

The COG was calculated as described in section 4.7.3 for the fricatives /s,

ʃ, ç, x/. Tongue positions were measured at the consonantal target position (cf. section 4.8.1). For /s/ and /ʃ/ the position of the tongue tip sensor was taken, for the other two fricatives the one of the tongue back sensor. The position of the jaw was measured at the articulatory target position as well.

Section 9.1 gives the results of the measurements of the COG. The section thereafter discusses the tongue position and section 9.3 the positions of the jaw.

9.1 Acoustic centre of gravity

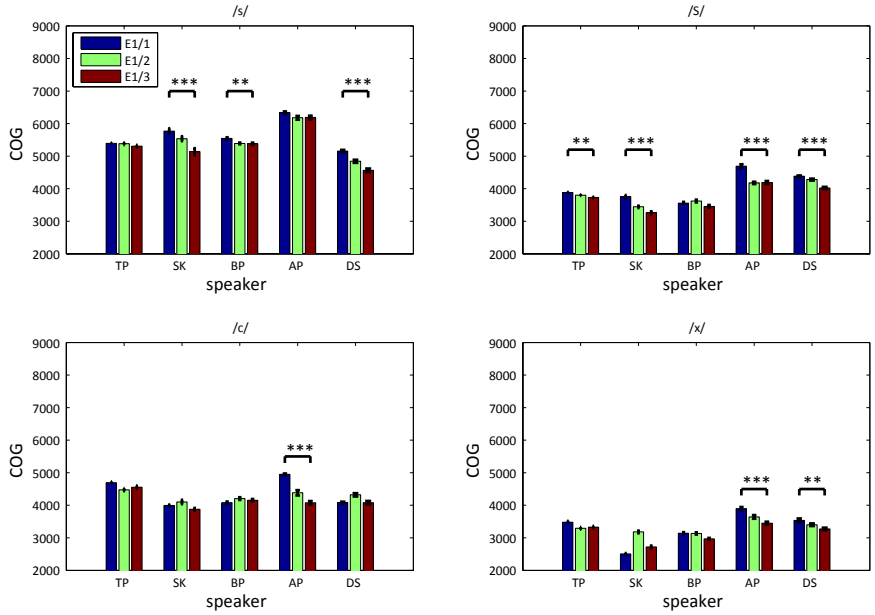


Figure 9.1: Mean COG values of four fricatives in the first three subsessions. Error bars show standard error. When the difference between sessions E1/1 and E1/3 is significant this is signalled by a bracket above the bars. Significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Significance levels are not shown when there is not a decrease from E1/1 to E1/3 but an increase.

Figure 9.1 shows mean COG values for each speaker for the four fricatives. Each subplot shows the results for one fricative. Each bar shows the mean over one session. Bars belonging to the same triple refer to one speaker and they are in the order of the recordings: The first bar corresponds to session E1/1 (pre-perturbed), the second to session E1/2 (perturbed with auditory

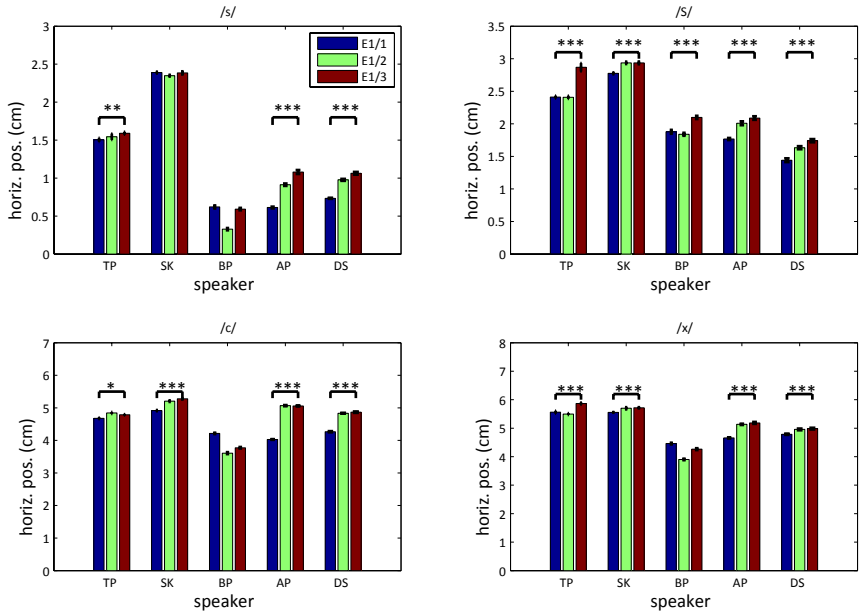


Figure 9.2: Mean horizontal position of the consonantal target position in four fricatives over the first three recordings. Higher values denote more retracted positions. Error bars show standard error. When the difference between sessions E1/1 and E1/3 is significant this is signalled by a bracket above the bars. Significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. If there is no increase from E1/1 to E1/3 the significance levels are not given.

feedback masking) and the third to session E1/3 (perturbed with auditory feedback available). The main tendency is that the COG is lower in session E1/2 than in E1/1 and that it decreases even more in E1/3. Exceptions are /ç/ and /x/ produced by speaker SK, /ʃ/ and /ç/ produced by speaker BP, /ç/ produced by speakers DS and TP. In general the lower COG value is thus not corrected when auditory feedback becomes available. On the contrary, the productions become "worse" in the sense of more unlike the ones in session E1/1.

The influence of the condition (unperturbed, perturbed with masking, perturbed without masking) was statistically assessed via a repeated measures ANOVA with data split by speaker and sound. Table A.8 in the appendix p.149, giving F-values and Greenhouse-Geisser corrected degrees of freedom, shows that the influence of the condition on the COG was in most cases significant. Tamhane T2-post hoc tests were calculated in order to compare sessions. When an F-value is printed in bold in the table this means

that all the differences are significant. When the F-value is printed in italics this means that the difference between session E1/1 and session E1/3 is significant. With one exception (/x/ by speaker SK), mean values decreased from an earlier session to a later one when a difference was significant.

9.2 Articulatory analysis: Tongue positions

Figure 9.2 shows the positional changes over sessions. In line with the decreasing COG-values there is a general increase in the x-values. This means that the place of constriction is more and more retracted. In the great majority of cases the values increase over all sessions, or they increase from one session to the following one and stay stable before or afterwards.

In only very few cases there is a tongue protrusion (lower x-values) in a perturbed session: Speaker BP has a protrusion for all sounds in the session without auditory feedback, but when auditory feedback becomes available he corrects for it.

Thus, all speakers except one retract their tongue at some stage, either immediately after perturbation onset or when auditory feedback becomes available. As the development of the COG, this result for the tongue shows that the retraction is not a mechanical one which is corrected when auditory feedback becomes available. Also, there is no reason to believe that the speakers search for a linguo-palatal contact pattern since (1) - even if this is not further discussed in this study - there is no difference between alveolar and central prosthesis and (2) there is no clear difference between sounds which are directly perturbed because their constriction is in the region of the artificial palate (/s, ʃ/) and other sounds. Table A.9 in the appendix, p.150 gives the results or the repeated measures ANOVA and Tamhane T2 post-hoc tests for the influence of the condition on the horizontal position. When a difference between session E1/1 and E1/3 is significant this is also shown in figure 9.3.

9.3 Jaw positions

Figure 9.3 shows the results for the jaw positions during /s/. For all speakers the jaw has a lower position in session E1/2 as compared to E1/1. The mean position in E1/3 is for all speakers a little higher than in E1/2. Tamhane T2 post-hoc tests show that the difference between session E1/2 and E1/3 is significant for all speakers.

For /ʃ/ (not shown here) the jaw becomes higher for only three speakers.

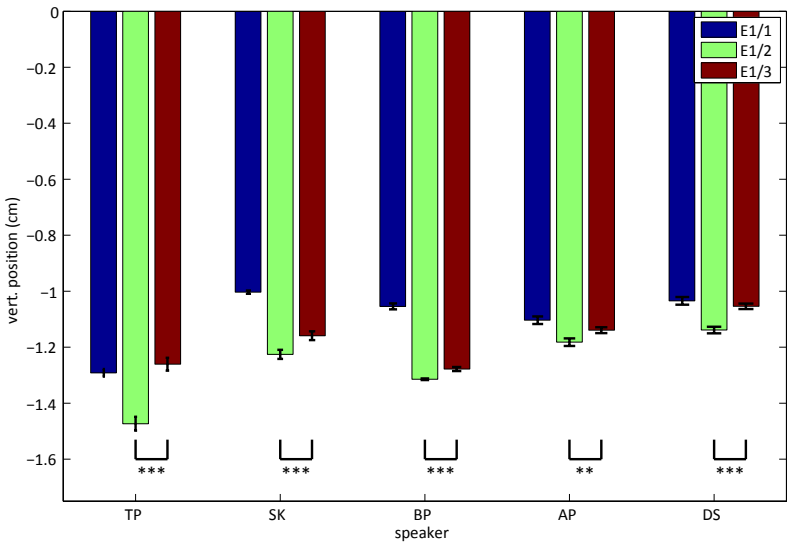


Figure 9.3: Mean vertical position of the jaw sensor at the consonantal target position in /s/. Higher values denote higher positions. Error bars show standard error.

For the palatal and the velar fricative no consistent pattern could be found. Table 9.1 gives the results of a repeated measures ANOVA for the influence of the session on the vertical position of the jaw sensor. The influence is always significant.

Table 9.1: Results of the repeated measures ANOVA for the influence of the condition on the vertical position of the jaw sensor with Greenhouse-Geisser corrected degrees of freedom. Significance levels: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Bold: The post-hoc tests resulted in a significant difference between sessions E1/2 and E1/3.

speaker	F
TP	**F(1.823, 19)=36.261*** **
SK	**F(1.357, 19)=59.547*** **
BP	**F(1.646, 19)=514.300*** **
AP	**F(1.728, 19)=8.845** **
DS	**F(1.562, 19)=21.806*** **

9.4 Conclusion

From the analysis in chapter 7 one could conclude that speakers manage to produce fricatives which can be correctly classified according to acoustic parameters already in the session where auditory feedback is masked. The present analysis has shown that auditory feedback, when it becomes available, does not necessarily contribute to an "improvement" of the productions in the sense that the COG becomes more similar to the one in the unperturbed session. In the majority of cases the centre of gravity is lower when no auditory feedback is available but it becomes even lower when auditory feedback is there.

Looking at the articulation one can see that the same tendency as in the acoustics can be found: When auditory feedback becomes available the productions do not become better, but they become "worse" in the sense that the articulatory position is even less similar to the original one when auditory feedback is available. The investigation of jaw positions has shown that the jaw is raised when auditory feedback becomes available. It is thus possible that the reason for retracting the tongue in /s/ is to direct the air jet against the lower incisors and thus give the sibilant the characteristic high amplitude noise. This acoustic characteristic seems to be more important than the adaptation of the centre of gravity.

Even though the high jaw position might be important for the sibilants, and the consequent retraction of the tongue can thus be explained, a retraction is not necessary for the palatal and the velar fricative. One could explain the change in these fricatives as a chain change - as has often been found in sound change. Chain changes aim at keeping the dispersion among the sounds: When one sound changes and this change makes the sound more similar to another sound, then this other sound moves "out of the way" ("push chain"). When it moves into the acoustic region of yet another sound, this third sound might also change.

In our case one can assume that the palatal and the velar fricative change because the sibilants have moved: Thus, /s/ was produced with acoustic characteristics which are too much like the ones of /ʃ/, so the postalveolar was retracted. This sound then became to similar to /ç/ so that /ç/ was retracted, and this change caused a positional change in the velar fricative.

The further development of the articulatory strategies is not discussed in this study. This is because there were great interspeaker differences which can be ascribed to different levels of effort, but also to different compensatory abilities. There are some speakers who manage to correct the lower COG values, others, however, keep the retracted positions.

Chapter 10

Motor equivalence in /u/

For rounded vowels like /u/ several articulatory strategies can produce the same acoustic output. Speakers could in principle use more or less lip rounding, and to compensate for that vary the tongue position. If one interprets the configuration of /u/ as two coupled Helmholtz resonators, F1 and F2 are the Helmholtz resonances from the back and the front cavity, respectively, which can be calculated according to the following formula:

$$F = \frac{c}{2\pi}\sqrt{\frac{A_c}{V * l}} \tag{10.1}$$

where c is the sound velocity, A_c is the area of the constriction, V is the volume of the cavity and l is the length of the cavity.

Retracting the constriction enlarges the front cavity and reduces the size of the back cavity. This results in a higher F1 (from the back cavity) and a lower F2 (from the front cavity). Tongue raising reduces A_c, the area of the constriction and thus lowers F1. Lip protrusion lowers both formants, F2 because the front cavity becomes longer and F1 because of the higher impedance the front cavity has on the back cavity. Thus, the same F1 and F2 can be produced by a number of articulatory configurations (Fant [1960]).

As discussed in chapter 1, it has already been shown that during several productions of /u/ some speakers vary in their application of tongue position and lip rounding. Three of the four speakers presented in Perkell et al. [1993] moved the tongue to a higher and more retracted position when they had little lip protrusion. When they had more lip protrusion they produced a wider constriction by lowering and fronting the tongue. For speakers who are confronted with a structural perturbation, these motor equivalent strategies could be useful. With an artificial palate speakers might at first lower their tongue position in order to avoid linguo-palatal contact in the still unknown environment and compensate for that by more lip protrusion. Later they

might use less lip protrusion and therefore a higher tongue position. The correlation between lip protrusion and constriction size can thus be expected to be very strong. On the other hand, if speakers make efficient use of motor equivalence, no correlation between either of the two articulatory parameters and one of the formants should be found.

10.1 Methods

Formant values were measured as described in section 4.7.2. Lip protrusion was measured by calculating the difference in the horizontal dimension between the upper lip sensor and the upper incisor sensor at the middle of the interval for which the formants were measured. By doing this it was assumed that the upper incisor sensor and the lip sensor were glued in about the same location on different recording days which seemed to be basically true.

The constriction size was measured as the shortest Euclidean distance between the palate and the tongue. In order to carry out this measurement, the palatal contour for each session was estimated by plotting all positional data which were recorded in the session. Then the palatal contour was estimated at the upper border of these measurements. For speakers BP and OP constriction measurements could not be carried out because the constriction was so much behind the end of the hard palate that the palatal contour could not be estimated with reasonable accuracy. In order to estimate the tongue contour, spline functions for the three tongue sensors were calculated for each production of /u/ at the acoustic target position. A comparison of the position of just the tongue back sensor (as was done in Perkell et al. [1993]) was not possible since the sensor position on the tongue varied between different experimental sessions. Afterwards, the constriction size was estimated as the smallest Euclidean distance between tongue contour and palate contour.

Bivariate Pearson correlations for the four parameters, lip protrusion, constriction size, F1 and F2 were calculated with SPSS 15.0.

10.2 Results

Table 10.1 gives the results of the correlations between lip protrusion, constriction size and the two acoustic parameters for the four speakers for whom significant correlations could be found. For speaker AP there was no significant correlation, so this speaker does not appear in the table. The first column gives the speaker, the second column shows the correlation coefficient for the correlation between lip protrusion and constriction size. If speakers

use motor equivalent strategies this correlation should be positive. The other columns give the correlations of the two articulatory parameters with F1 and F2. If speakers use motor equivalent strategies effectively in order to keep the acoustic output constant, there should be no correlations between the articulatory and the acoustic parameters.

Table 10.1: Correlation coefficients and significance level of the correlation between lip protrusion and constriction size (second column), lip protrusion and F1 (third column), lip protrusion and F2 (fourth column), constriction size and F1 (fifth column), constriction size and F2 (sixth column). Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

speaker	lip-const	lip-F1	lip-F2	const-F1	const-F2
DS	0.568***	-0.126	-0.220*	-0.173	-0.033
KD	0.537***	-0.330**	0.065	-0.197	-0.108
SK	0.660***	0.006	0.118	0.037	0.093
TP	0.235*	-0.351***	-0.313**	-0.014	0.042

Thus, a significant positive correlation between lip protrusion and constriction size could be found for four speakers. For three speakers (KD, DS and TP) there are also significant correlations between lip protrusion and one or both of the formants. These could be seen as cases of insufficient compensation via tongue positioning: There is too much lip protrusion which is not completely compensated for.

In order to show more details the results for lip protrusion vs. constriction size for speaker DS are shown in figure 10.1. Different markers represent different sessions. The figure is meant to illustrate two points which could be seen for all the four speakers.

First, with regard to the single sessions, an important point is that the correlation is weaker or non-existent within sessions. The correlation is therefore a result of the articulatory behaviour across sessions. Within each session the speaker has a preferred strategy, a preferred constriction size and a preferred degree of lip protrusion. In session E1/1 (squares), for example the lips are only slightly protruded and the constriction is rather small. In the last perturbed session (diamonds) there is a lot of lip protrusion, and the constriction is larger.

The second point to make is that it is in general not the case that the speaker at perturbation onset ('x'es and asterisks) has more lip protrusion and a greater constriction than later in the adaptation time. In contrast to what was hypothesised at the beginning the speakers in general do not use wider constrictions at perturbation onset.

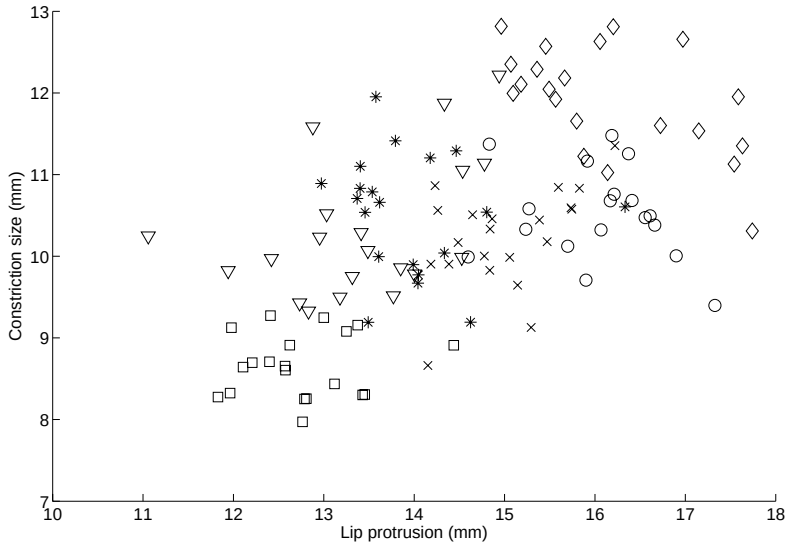


Figure 10.1: Lip protrusion and constriction size measurements for speaker DS. Different markers represent different sessions: E1/1: squares, E1/2: asterisks, E1/3: x, E2/1: triangles, E3/1: diamonds, E3/2: circles.

10.3 Conclusion

The investigation presented here has shown that evidence for motor equivalent strategies in /u/ production can be found for the majority of speakers. These speakers have more lip protrusion when the constriction is large and less lip protrusion when the constriction is small. For one speaker the motor equivalent strategies result in a stable acoustic output (there is no significant correlation between the articulatory and the acoustic parameters). For three further speakers the motor equivalent strategies are not completely successful. They do not lower the tongue enough for the high degree of lip rounding found. They thus have a motor equivalence strategy which is just not efficient enough. For one speaker no correlations were found, for two speakers the measurement could not be carried out because the constriction was in the region of the soft palate which could not be estimated reliably.

Contrary to an expectation stated in the beginning, speakers do not consistently use more lip protrusion in the early perturbed sessions in order to compensate for an articulatorily easier lower tongue position. Rather, the strategies vary over the sessions. Each speaker has a certain preference for an articulatory strategy in each session. This supports the assumption that speakers try out different strategies over the adaptation time in order to find

an optimal one. The correlations found here are a little stronger (higher correlation coefficients) than the ones presented in the study by Perkell et al. [1993]. A reason for that could be that, in contrast to this earlier study, in the present study the speakers are confronted with a perturbation. They might thus be more inclined to try out several articulatory strategies. Judging from the sensor plots presented in Perkell et al. [1993], lip protrusion varies in all cases considerably less than a centimetre in the horizontal dimension. The ranges for the speakers presented here are between 7 and 12 mm.

The missing correlations between constriction size and lip protrusion for one speaker could be due to methodological problems since it is difficult to measure constriction size. Another possibility could be that the speaker does not try to improve the productions any further, and the variability in constriction size and lip protrusion is thus too low. This could be because the productions are acoustically acceptable and efficiently executed. A third possibility could be that the speaker is not sensitive enough to the small acoustic changes due to differences in articulation and therefore does not try to improve the adaptation (Perkell et al. [2004]).

The use of these motor equivalent strategies supports acoustic perceptual primitives since the articulatory strategies vary whereas the acoustics stay constant.

Chapter 11

Movement optimisation

The results discussed so far have shown that the acoustic targets of the vowels are achieved rather soon after perturbation onset whereas speakers need more time in order to adapt the fricatives. A further finding is that even when the acoustic targets are reached, speakers go on to change their articulation for reasons that are often not obviously linked with acoustic or articulatory requirements.[1] A third aspect from the discussion of vowel acoustics is that speakers show a rather low degree of variability within a session: Compared to the variability found across sessions (defining the *regions* in chapter 5) the session specific *subregions* in acoustic space are rather small.

The second finding, that speakers still change their productions even after the acoustic targets have been reached, suggests that apart from reaching an acoustic target, speakers might have a second aim. As discussed in chapter 2, there is experimental evidence that articulator movements are organised in a certain way in order to minimise the articulatory effort. Speakers' second aim during adaptation could therefore be to reduce the articulatory effort and thus find an optimal mode of production. The aim of the analysis presented here[2] is therefore to investigate whether a movement optimisation based on jerk minimisation could explain the observations made for vowel acoustics.

As discussed in section 2, speech production has been assumed to involve an internal model of speech production in the speaker's brain in which mappings of motor commands to acoustic outputs are stored. During speech production, a subcomponent of this internal model (the forward model) predicts the acoustic output and the kinematic properties for a motor input so that for a given motor input the speaker knows what he is going to produce (acoustically and articulatorily) before he actually produces the sound.

[1] But see the exception discussed in chapter 6: /i/ and /e/ seem to be fronted in order to increase the acoustic or articulatory distance.

[2] Preliminary results of this analysis can be found in Brunner et al. [2007]

Furthermore, another subcomponent, the inverse model, proposes a certain motor input for a desired acoustic output so that the speaker can infer what to do from the objective he wants to reach in the output domain. Since the mapping *motor input-acoustic output* is a many to one mapping, a selection process must be involved. In some approaches (e.g. Jordan [1996]) it is assumed that the most optimal movement, in terms of several parameters, among which there is the smoothness of the movement, is selected in this process.

Following this model, since the speakers in this study have completed speech learning, they can be expected to have a well trained internal model in the pre-perturbed session which predicts the correct acoustic output from a motor input together with the kinematic characteristics of the movement, but also provides a motor input for a desired output leading to a movement which involves minimal articulatory effort (e.g. a minimal jerk trajectory).

When speakers are first confronted with the palatal prosthesis, one could think of two different scenarios: In the first scenario speakers could change the global parameters of the articulation without changing the kinematic properties of their articulatory movements as such. For example, experiencing that the palate is on the whole lower, they might open the jaw a bit more. This will lead to a start and end position which is different in absolute spatial terms, but it will in general not lead to a different movement organisation which is measurable in differences in peak velocity, temporal location of the velocity peak, movement amplitude and articulatory effort.[3] During this phase, when there is no reorganisation of the movement but only a reparametrisation (cf. chapter 3) the articulatory effort would stay the same as in the unperturbed session.

The second scenario: Either immediately after perturbation onset or after the just mentioned "shift" phase speakers might start to reorganise their movements in order to improve the acoustic output or maybe to reach an articulatory target. For example, they might change the tongue bunching. This should result in an increase in articulatory effort. In both scenarios, whether there is a reorganisation or not, the forward model is retrained during the adaptation phase, since new example pairs *motor input-acoustic output* are presented to it but only in the second case the kinematic information should differ.

If speakers ever leave the reparametrisation phase and the kinematic prop-

[3]This holds if one ignores possible changes in articulator stiffness which result from a wider opening of the jaw (cf. Shiller et al. [2002]) which would result in different kinematic properties. Support for this assumption comes from Laboissière et al. [1996] who found for jaw movements that local motor command variations induce the same articulatory changes whatever the starting position.

erties of their movements thus change, at some point, when there is a sufficient number of new *input-output* examples in the internal model, the inverse model should start to select movements which involve less articulatory effort. This should result in a drop of the articulatory effort of the movement.

Consequently, for the evolution of the articulatory effort after the insertion of the palate one can expect that it at first increases (after a possible stable phase) and later drops. The first phase, when the articulatory effort increases, will be called *training phase* because during this phase the internal model is retrained, and the second phase, when effort decreases, will be called *optimisation phase* since during this phase optimal articulatory movements are selected.

The more subtle properties of this increase-decrease pattern should depend on a number of factors. The *point in time of the optimisation onset* (the session with maximal articulatory effort) could depend on the degree of difficulty of the movement. For example, a movement towards a vowel should in general be easier than a movement towards a fricative because the articulatory target of the fricative requires a high degree of precision. If it is speakers' first aim to produce a certain sound, and movement optimisation is only secondary, speakers can be expected to start to optimise their movements earlier in vocalic gestures than in consonantal gestures. Another possibility is that movement optimisation starts at the same time for all sounds. This could mean that movement optimisation is not speakers' secondary aim but that it has equal importance as reaching the phonemic target. Yet a third possibility is again based on the internal model approach discussed in Jordan [1996]: Movement optimisation is a process which is completely independent from the fulfillment of the task. The task can be correctly fulfilled when the forward model, which predicts the acoustic output of a motor input and the kinematic properties, is correct. Movement optimisation, however, is independent of this and selects more optimal movements among the input-output pairs which happen to be in the internal model.

The effort pattern might furthermore depend on *the importance of smooth movements for the acoustic output.* It is possible that smooth movements are sometimes necessary in order to produce a certain acoustic output. For example, for the correct production of /z/ it is necessary that the tongue tip reduces its velocity sufficiently early before reaching the palate because otherwise a closure instead of a constriction will be produced. A smooth movement is therefore not only a matter of articulatory effort but it is vital for the production of the sound. For the acoustic output of /t/, on the other hand, it is not absolutely necessary to have a smooth movement. The tongue might reach the palate with maximal velocity and this will hardly influence the acoustics. Consequently, it is possible that the speaker starts

to produce optimised movements earlier for the gesture towards the fricative than during the gesture towards the stop. In this case it would not be optimisation in terms of effort reduction what is measured. It is rather that effort reduction and closure prevention might result in the same measurable parameter (minimal jerk trajectories).

Yet another aspect concerning effort development is whether *the involvement of the articulator in fulfilling the task* plays a role. It is possible that movements of an articulator which is involved in the production of a sound are optimised differently than movements of other articulators in the same gesture. For example, when the correct production of the constriction in /i/ makes difficulties, the optimisation of the movement measured for the tongue dorsum sensor might start later than the optimisation of the movement measured at the tongue tip because the latter articulator is not directly involved and can therefore move freely.

A final question is whether *biomechanical factors* are involved. It is possible that some articulators are more apt for movement optimisation than others. If the internal model involves a generalisation over input-output pairs it might be that the movements of more "solid" articulators as the jaw are easier to optimise than flexible articulators such as the tongue. Thus, it could be that the pattern is found more consistently for the jaw than for the tongue.

In order to shed light on all these questions, in the present chapter movement optimisation in two vowels, /o/ and /i/ and two consonants /t/ and /z/ will be compared. Three articulators are investigated: tongue tip, tongue dorsum and jaw. Related to the aspects mentioned just above: The sounds present different degrees of difficulty (/o/ and /t/ are easy to adapt, the others are a bit more difficult). Second, /z/ necessitates a certain critical constriction whereas the stop does not. Third, in vowel production different articulators are involved than in the production of the alveolar consonants. Fourth, there are three articulators which can be compared so that one can see whether biomechanical factors play a role.

11.1 Methods

Consonantal gestures leading towards /z/ and /t/, and vocalic gestures leading towards /o/ and /i/ were measured for the tongue tip, tongue dorsum and the jaw as described in section 4.8.2. In order to assess articulatory effort, the jerk for the gestures was measured according to the formula given

in Nelson [1983]:

$$J = \frac{1}{2}\int_0^T \dot{a}^2(t)dt \tag{11.1}$$

As can be seen from the formula, all other things being equal J is higher for longer segments and lower for shorter ones. In order to account for the differences in duration for different sessions J was divided by the number of samples. The resulting value, a temporally normalised tangential jerk, will be referred to simply as *jerk*. Means, standard deviations and standard errors for each session were calculated.

Following the discussion just above, the expected *increase-decrease* pattern, if it exists, would be marked by three points in time: (1) session E1/1 where maximally optimised movements and thus low jerk values can be expected, (2) a perturbed session in between sessions E1/1 and E3/1 where the jerk is maximal and (3) session E3/1 where, provided that the adaptation time was sufficiently long, again maximally optimised movements can be expected. The phase from E1/1 to the session with maximal jerk is, according to the definition given above, the *training phase*, and the phase from the session with maximal jerk to session E3/1 is called the *optimisation phase*. For all the cases where an increase followed by a decrease could be found, repeated measures ANOVAs and Tamhane T2 post-hoc tests were calculated in order to compare these three sessions.

11.2 Results

Results were thus gained for seven speakers, four sounds and three articulators. As an introduction to the kind of results gained from the analysis, the results for the tongue tip of /o/ will be discussed. Afterwards, the classification into speakers, sounds and articulators will be given up in favour of a grouping according to the pattern found. Three patterns will be distinguished:

- *increase-decrease*. The jerk increases from session E1/1 to a higher value at some session in between E1/1 and E3/1 and then decreases until session E3/1. Both the difference in jerk between E1/1 and the session with maximal jerk, and the one between this session and E3/1 are significant. There is thus a training phase and an optimisation phase.

- *increase only*. The jerk increases from session E1/1 to a higher value at some later session. The difference in jerk between E1/1 and this

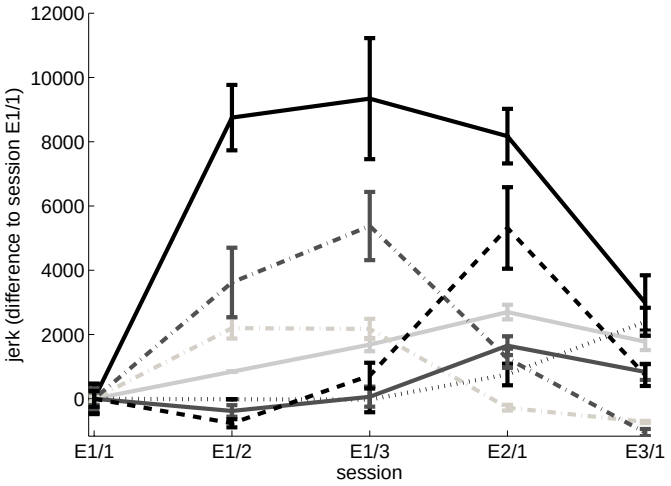


Figure 11.1: Differences in jerk to session E1/1 during the tongue tip gesture towards /o/. Different grey shades and linestyles correspond to different speakers. TP: solid black, KD: solid light grey, OP: dotted black, SK: dash-dotted light grey, BP: solid dark grey, DS: dash-dotted dark grey, AP: dashed black. Error bars show standard error.

session is statistically significant. No significant decrease can be found in E3/1 as compared to any of the preceding sessions.

- *decrease only.* The jerk decreases either directly from session E1/1 or from a later session with maximal jerk until session E3/1. This decrease is significant. If there is an initial increase it is not statistically significant.

As announced, for introductory purposes, an example for the results gained is shown in figure 11.1. The figure shows the development of the jerk in the tongue tip gesture towards /o/. Each line given in the figure shows the results for one speaker. The abscissa gives the sessions. The values shown in this figure are differences between the mean jerk values calculated for a session and the unperturbed session E1/1, consequently, the mean value of session E1/1 is 0. The reason for showing differences rather than absolute means is purely presentational. The absolute values differ a lot for the speakers so that more information can be gained from the figure when the means are normalised by the unperturbed session. Standard errors (given by the error bars) refer to the absolute values. Even in this presentation with the thus normalised values one can see that the jerk is very speaker specific, some speakers have very high differences (e.g. TP, solid black), some have very

Table 11.1: Cases in which the pattern *increase-decrease* was found ($p < 0.05$). Column 1: speaker, column 2: sound, column 3: sensor, column 4: session with maximal jerk value

speaker	sound	sensor	max.	speaker	sound	sensor	max.
DS	/o/	ttip	E1/3	AP	/i/	jaw	E2/1
KD	/o/	ttip	E2/1	AP	/t/	ttip	E2/1
SK	/o/	ttip	E1/2	DS	/t/	ttip	E1/3
AP	/o/	ttip	E2/1	KD	/t/	ttip	E2/1
TP	/o/	tdor	E2/1	TP	/t/	ttip	E1/3
DS	/o/	jaw	E2/1	AP	/t/	tdor	E2/1
SK	/o/	jaw	E1/2	DS	/t/	tdor	E2/1
TP	/o/	jaw	E1/2	SK	/t/	tdor	E1/2
AP	/z/	ttip	E1/3	AP	/t/	jaw	E2/1
TP	/z/	ttip	E1/2	DS	/t/	jaw	E2/1
AP	/z/	tdor	E1/32				

low ones (e.g. OP, dotted black).

For six of the seven speakers (all except OP in dotted black) the *increase-decrease* pattern can be found. As will be shown in detail later, for all six speakers the increase (the difference between session E1/1 and the session with maximal jerk) is significant. For five of these six speakers (all except BP, solid dark grey) the decrease (the difference between the session with maximal jerk and session E3/1) is significant. Except for two speakers (AP, dashed black, and BP, solid dark grey) the increase is monotonic. The development in the early sessions for speakers AP and BP might present cases in which there was no reorganisation at perturbation onset but where it started later. In summary, six speakers show the pattern *increase-decrease* and two speakers (OP, dotted black and BP, solid dark grey) show the pattern *increase only*.

After this exemplary description of the results, the discussion will turn to the patterns which can be found for other sounds and articulators. Table 11.1 gives the cases for which the first pattern, *increase-decrease* was found. More detailed results (with means for all sessions and standard deviations) can be found in the appendix (table A.12, p.153). The tables presented here in the text do not give very detailed information. The reason for including them is rather to show how often a pattern was found and for which articulators and sounds it was found.

As discussed above already, the *increase-decrease* pattern can be found for the tongue tip gesture towards /o/. Apart from that it can be found for

Table 11.2: Cases in which the pattern *increase only* was found ($p < 0.05$). Column 1: speaker, column 2: sound, column 3: sensor, column 4: session with maximal jerk value

speaker	sound	sensor	max.	speaker	sound	sensor	max.
BP	/o/	ttip	E2/1	KD	/i/	ttip	E2/1
OP	/o/	ttip	E3/1	TP	/i/	tdor	E2/1
AP	/o/	jaw	E2/1	DS	/i/	tdor	E2/1
BP	/o/	tdor	E2/1	KD	/i/	tdor	E3/1
KD	/o/	jaw	E3/1	TP	/i/	jaw	E1/2
DS	/z/	ttip	E2/1	DS	/i/	jaw	E2/1
OP	/z/	ttip	E1/3	KD	/i/	jaw	E2/1
SK	/z/	ttip	E1/2	SK	/t/	ttip	E1/2
DS	/z/	tdor	E3/1	OP	/t/	ttip	E1/3
KD	/z/	tdor	E2/1	KD	/t/	tdor	E2/1
DS	/z/	jaw	E2/1	OP	/t/	tdor	E1/3
KD	/z/	jaw	E2/1	TP	/t/	jaw	E1/2
TP	/z/	jaw	E3/1	KD	/t/	jaw	E3/1

the jaw (three speakers) and once also for the tongue dorsum. Furthermore, it can be found for the tongue tip gesture towards /t/ (four cases), the tongue dorsum gesture (three cases) and the jaw gesture (two cases). For /i/ and /z/ the pattern is rare. In general, the *increase-decrease* occurs most often for /o/ and /t/ and for the tongue tip movement.

With respect to the optimisation onset, the session with maximal jerk (column 4 in the table) one cannot find a pattern. It is not that optimisation starts earlier for easier sounds than for more difficult sounds, and it is not the case that speakers show a certain preference, nor that optimisation starts earlier or later for a certain articulator.

The second pattern *increase only* (cf. table 11.2) was found about equally often for all the four sounds. Similarly, there is no clear preference for a certain articulator. In 10 cases jaw movements are involved, but in 8 cases tongue tip or tongue dorsum movements are involved.

The pattern *decrease only* was rather rare as compared to the first two patterns (table 11.3). The pattern occurs most often for the tongue dorsum. For all cases which are not listed in one of the three tables none of the three patterns was found.

Table 11.3: Cases in which the pattern *decrease only* was found ($p < 0.05$). Column 1: speaker, column 2: sound, column 3: sensor, column 4: session with maximal jerk value.

speaker	sound	sensor	max.
SK	/o/	tdor	E1/1
OP	/o/	jaw	E1/1
BP	/i/	ttip	E2/1
DS	/i/	ttip	E1/2
BP	/i/	tdor	E1/1
BP	/t/	tdor	E2/1
TP	/t/	tdor	E1/3
BP	/z/	ttip	E1/3
TP	/z/	tdor	E1/1

11.3 Discussion

Even if the results are somewhat inconsistent, the analysis presented here suggests that movement optimisation exists during the adaptation towards a perturbation as the one described here. To summarise the main results:

- The pattern *increase-decrease* can be found most often for the easiest sound, namely /o/.
- The pattern *increase-decrease* can be found most often for the tongue tip sensor.
- The pattern *increase only* can be found about equally often for all sounds and all articulators.
- The pattern *decrease only* is rare.
- The point in time of optimisation onset varies inconsistently.

The fact that the pattern *increase-decrease* can be found at all supports the assumption that there is indeed a retraining of the internal model. During the first phase new example pairs of motor input and acoustic output are presented to the model, and during the second phase more and more optimal movements are selected.

The second pattern *increase only* can be explained if one assumes that the adaptation time was too short. Maybe not enough example pairs have been provided to the internal model in these cases so that it was not yet possible to start optimisation.

The third pattern is hard to explain, except if one assumes that the speakers have got used to not only the palatal prosthesis but also the experimental method. In session E1/1 they were still inexperienced with EMA but later on they produced more and more optimal movements under EMA. However, since this pattern is rare one can assume that the adaptation towards the prosthesis had a stronger influence on the articulatory behaviour than the adaptation towards the experimental method.

In general, the *increase-decrease* pattern occurs more often for the easy sound /t/ than for the difficult sound /z/. This is somehow astonishing since smooth movements are actually acoustically necessary for /z/ and visual inspection of the trajectories suggested that the movements in /z/ are much smoother than in the stop. Looking at the absolute values for the fricative, however, this can be explained. In fact, the jerk values are much lower for the fricative and the reason for that is most likely the low velocity in this sound. The jerk as a parameter measures not only articulatory effort but, and this seems to be crucial here, it is heavily dependent on velocity. Since the velocity in /z/ is so much lower than in /t/, one might just not be able to see the movement optimisation effect for /z/ because the signal to noise distance is too small.

Furthermore, the *increase-decrease* pattern occurs more often for the back vowel /o/ than for the front vowel /i/. For the front vowel the *increase-only* pattern is predominant. Here, one explanation could be that /o/ is less perturbed acoustically and articulatorily.[4] Speakers can therefore concentrate on optimisation because the acoustic output is soon acceptable. Additionally, due to the availability of motor equivalent strategies (tongue raising versus lip protrusion versus larynx lowering) there are very many possibilities to produce /o/ so that there is a great number of possibilities among which the most optimal one can be chosen.

The fact that the *increase-decrease* pattern can be found more often for the tongue tip sensor than for other sensors suggests that optimisation does not depend on the involvement of an articulator in the production of a sound as was hypothesised in the beginning. If this was the case the tongue dorsum should show less optimisation for the vowels and the tongue tip should show less optimisation for the consonants. What can be found, however, is that the tongue tip for both sound classes shows a clearer optimisation than the other articulators. This could again be due to the generally higher velocity of the tongue tip which allows for more reliable measurements.

[4]Support for this hypothesis is given by results for other easy sounds such as /u/ and /y/ (cf. Brunner et al. [2007]).

The fact that the optimisation onset does not vary consistently for sounds suggests that movement optimisation is at least in part independent of the acoustic output. Comparing the two vowels, even if the front vowel was more perturbed and more difficult to adapt, the optimisation onset was not consistently later. This suggests that movement optimisation is a process which can start even if the forward model is not yet correct.

Looking at the absolute results of the jerk measurements given in the appendix one can notice a very high speaker dependence, but also fundamental intraspeaker variability. This shows that measurements of this kind are very difficult. There are obviously several mechanisms involved which result in the same measureable parameter, the jerk, for example the velocity of the movement and the movement amplitude, which all contribute to what might be called articulatory effort, but which make it difficult to assess in which cases there is movement optimisation in the sense in which it is used in the internal model approach.

To conclude, even if it is difficult to measure, optimisation seems to exist in the adaptation process towards a perturbation. This finding explains why there are changes in articulation even when the acoustic output is correct and these changes are not directed towards a clear articulatory target. It is because speakers try to find more optimal articulatory strategies. The finding furthermore explains why speakers show so little articulatory variability especially towards the end of the experiment. Speakers select certain optimal strategies and this reduces the overall variability.

Chapter 12

General discussion

12.1 Summary of the results

The aim of the study presented here was to investigate whether confronted with an articulatory perturbation speakers adapt towards acoustic or articulatory targets. The expectation was that from the results of the experiment new insights into the nature of speech production tasks and possibly also phonemic perceptual primitives could be gained. Speech production tasks were assumed to consist of two components, (1) the acoustic or articulatory perceptual target which is transmitted to the listener, and (2) a motor representation which facilitates the articulation. More concretely, the phoneme representation which the speaker has saved in his brain consists of an acoustic "image" of the sound, maybe a formant pattern, or possibly an articulatory "image", maybe an articulatory gesture. This representation of the sound is the same for the speaker and the listener and it enables the listener to understand what the speaker says. Additionally, the speaker has, for himself a motor representation which helps him to articulate the sound. If the perceptual representation is acoustic, this motor representation enables the speaker to produce speech even when auditory feedback is not available. In any case, no matter whether the perceptual representations are articulatory or acoustic, this motor representation should help the speaker to produce articulatory movements which involve little effort.

When speakers are confronted with an articulatory perturbation they should have two aims, first, to reach the perceptual targets and second, to produce optimal movements. The present study investigated whether during the adaptation time evidence can be found for a compensation towards articulatory or acoustic perceptual targets, and to what extent speakers use motor representations.

In order to shed light on this question a perturbation experiment with seven speakers was carried out. Speakers vocal tract shape was modified by a palatal prosthesis which was worn by the speakers for two weeks. Over this period, speakers were recorded several times via electromagnetic articulography and acoustically.

The first analysis investigated a premise for speech production tasks in the acoustic domain (which however as such does not present evidence that they really are in the acoustic domain), namely that all sounds can be distinguished unambiguously by acoustic characteristics. As shown by the discussion of formant values presented in chapter 5, for each speaker separately, there are preferred regions in acoustic space defined by F1, F2, F3 and duration, which, over sessions, overlap. Splitting data by speaker and session, however, there is hardly any overlap among tense vowels. Including lax vowels in the discussion, however, the vowels can no longer be distinguished from each other within a session. Thus, other mechanisms, such as the formant contour, f0 or the position of the sound in the syllable must be involved. A discriminant analysis carried out for several acoustic parameters supported the result which was gained from the formant analysis: Nearly all productions of tense vowels could be assigned to the correct phoneme class. However, comparing different sessions, the classification was less good in the first perturbed session when auditory feedback was masked and even less well separated in the session thereafter when auditory feedback was available. From this session onwards, however, speakers quickly returned to the initial high classification values.

A comparison of the formant values over time showed that they vary in a way which in general cannot be interpreted in terms of an improvement of the acoustic output: It is not possible to see a clear direction or a development towards the unperturbed values. Also, apart from two exceptions, it is not possible to see a clear development of the formant values which could be interpreted in articulatory terms. For example, it is in general not the case that F1 or F2 rise or fall consistently over sessions so that one could assume that speakers are producing positions which are maybe lower and lower or more and more retracted.

In chapter 6 the two exceptions were investigated where a consistent rise in F2 was found over the adaptation time. This was the case for the /i/ and /e/ productions of the majority of speakers. In the analysis presented in chapter 6 articulation and acoustics were investigated at the same time in order to see whether a development exists in both domains or whether it is linear in the acoustic domain only. This would have supported perceptual primitives in the acoustic domain. However, a linear development could be found in both domains and it is assumed that its aim is to create greater

differences between the vowels of the German vowel inventory.

The analysis described in chapter 7 had similar aims as the one in chapter 5, but this time for the fricatives. The first aim was to find separate regions in acoustic space for each phoneme, and the second aim was to investigate a possible development towards the acoustic characteristics of the unperturbed productions. The results show that, within a session and for each speaker separately, it is possible to classify most of the fricative productions by purely acoustic information except for the voiced and the voiceless alveolar fricative. This is because the spectral characteristics were calculated for the frequency range of 700 Hz to 12 kHz so that the voicing characteristics distinguishing the two sounds were excluded. In contrast to the vowels, the classification scores did not change considerably over the adaptation time.

Similarly to the vowels, the fricative productions still change, even at a stage of the adaptation where the productions can be correctly classified. In contrast to the vowels, however, the parameters do not vary randomly, rather there is a clear improvement of the productions in the sense that the acoustics develops towards the values of the unperturbed session. This development, however, cannot be assigned to a single acoustic parameter. Rather, the acoustic parameters used to characterise a fricative vary.

Chapter 8 returned to the vowels and investigated the influence of auditory feedback by comparing F1 of the vowels in session E1/2, where auditory feedback was masked with session E1/3 where auditory feedback was available. The results show that the productions improve in the session with auditory feedback available (they become more similar to the productions in the unperturbed session). One result of the analysis is that the productions of vowels with more linguo-palatal contact are in general worse in both sessions than the ones of the vowels with less linguo-palatal contact. This is probably due to the fact that high vowels are more difficult to adapt than low vowels because they do not allow for much articulatory variability. A second finding is that the improvement is greater for vowels with less linguo-palatal contact than for vowels with much linguo-palatal contact. Thus the already better low vowels improve more when auditory feedback becomes available. These two findings could be seen in relation: Improvements are more difficult for high vowels, since they are in general more difficult to adapt. On the other hand, however, one could argue that the strong improvement for the vowels with less contact is the result of auditory feedback. When auditory feedback is not available the sounds with lots of contact can easily be adapted with the help of tactile feedback. This is more difficult for the other sounds since tactile feedback cannot be used up to the same degree. When auditory feedback becomes available, speakers can make intensive use of it to improve the sounds with less linguo-palatal contact.

Chapter 9 returns to the fricatives and, as chapter 8 did for the vowels it compares the productions in the session with auditory feedback masked and with auditory feedback available. In contrast to the vowels where an improvement of the formant values can be found in the sense that the formant values became more similar to the ones from the unperturbed session, the fricative productions seem to become "worse". Acoustically, the COG values decrease over the three sessions, from the unperturbed session over the session without auditory feedback to the session with auditory feedback available. Correspondingly, for tongue positions it was found that speakers more and more retract the tongue. As a reason for this retraction it was hypothesised that only with a retracted tongue position are speakers able to keep a jaw position which is high enough so that during the sibilants the air jet can be directed towards the lower teeth. In order to test this hypothesis the jaw position was investigated and it was found that the jaw is indeed higher in the session when auditory feedback is available than in the session when it is not. As a reason for why the articulatory position of the palatal and the velar stop are retracted as well even if a high position of the jaw is not necessary for these sounds it is proposed that speakers change the positions of all fricatives in order to keep the articulatory or acoustic distance between the phonemes.

Chapter 10 presents a rather classical test for acoustic perceptual primitives by investigating motor equivalent strategies during the production of /u/. This sound is usually produced with a double constriction, one at the lips and one in the velar region. When the lips are protruded a bit more and the constriction in the back of the mouth becomes wider the formant values stay constant. Equally, when the lips are less protruded and the tongue is raised, the acoustic characteristics of the vowel still stay the same. For the majority of our speakers such motor equivalent strategies could indeed be found. There is a correlation between lip protrusion and the size of the constriction in the back of the mouth. This result shows that speakers might use more than one articulatory strategy in order to produce a certain acoustic output. Thus, the acoustic output stays stable while the articulation varies.

Chapter 11 investigates whether speakers optimise their movements over time. For the movement of the tongue tip in the articulatory gesture towards /o/ it was found that after perturbation onset the articulatory effort, measured as tangential jerk over the gesture, at first increases and later decreases. For other articulators the same result could not be found, and even for the tongue tip gesture during different sounds the results were not clear. It is supposed that this could be due to the many parameters which can influence the jerk measurement (velocity of the articulator, movement amplitude).

12.2 Discussion

The results presented here suggest that both - articulatory and acoustic targets - exist in the speech production tasks. A number of articulatory components, however, seem to exist for reasons of motor control only. An attempt will be made to separate these from the perceptual primitives which are transmitted to the speaker.

12.2.1 Adaptation toward articulatory or acoustic targets

A precondition for invariant acoustic targets involves separate regions in acoustic space for each phoneme. The results presented here show that - within limits - these regions exist. They do not exist in absolute terms but they vary from session to session. Thus, as has been found in several studies already, acoustic invariants as proposed by the *Theory of Acoustic Invariance* do not exist. Rather, the results support the *Adaptive Variability Theory*: The acoustic targets vary in dependence on the "communicative situation" and what is important is that phonemes can be separated within the same situation. Support for this can be seen in the rather high classification scores for the vowels and fricatives when each session was analysed separately. Further support comes from the chain shift observed for the fricatives. Probably in order to reach a certain acoustic target one of the sounds (/s/) had to be produced with a retracted tongue. In order to retain the space between the fricatives of the German phoneme inventory, the other fricatives were produced more retracted as well. This result is consistent with the one from the bite block study presented in McFarland and Baum [1995]. In this experiment speakers shifted the COG in /p, t, k, s, ʃ/.

The adjustment of the jaw which could be found for the speakers in the present study when auditory feedback becomes available, suggests that speakers change their articulation in order to produce sibilants with the important acoustic characteristic of high frequency noise. Without prosthesis the air jet was directed against the upper teeth. When the prosthesis is inserted this is no longer possible. Without auditory feedback speakers more or less keep the original way to articulate the sound, but when auditory feedback becomes available they change their articulatory pattern in order to reach a certain acoustic pattern. Here the articulatory (possibly motor) representation is overrun by the acoustic representation.

Another case where the articulatory representation was overrun by the acoustics are the motor equivalent strategies found in the production of /u/. Here the articulation varies but the acoustic output stays constant. This

can only be explained if one assumes an acoustic phoneme representation. Speakers' aim seems to be to produce a certain acoustic output and therefore it is this acoustic output which stays stable.

Further support for acoustic representations comes from the development of the fricative productions over the adaptation time. As shown by means of a discriminant analysis, the productions become acoustically more and more similar to the unperturbed productions. This development is difficult to allocate to a single acoustic parameter, as for example the COG, but it is equally difficult to allocate it to a specific articulatory parameter.

A last piece of evidence comes from the different behaviour of /t/ and /z/ which was found rather along the way when movement optimisation was investigated. These two sounds share a number of articulatory properties, but they are acoustically very different. Under perturbation speakers reduced the velocity in /z/ but they did not do so in /t/. If speakers had an articulatory configuration in mind they should have adjusted both velocities in order to fit the new surroundings. Acoustically, however, the adjustment in /t/ is not necessary since it does not influence the acoustic output considerably.

An argument, especially from Direct Realists, against much of the evidence for acoustic perceptual primitives is that the acoustic signal does contain all this information - there are separate regions for phonemes in acoustic space - because the recovery of the articulatory information from this signal is only possible when the acoustic information provides it. Thus, one could say that the separate regions in acoustic space found for the vowels and the fricatives are there in order to be able to say, for example, whether it was an alveolar constriction or a velar closure which the speaker produced. Similarly, the productions must become more dispersed in the acoustic space because it allows for a better recognition of the articulatory gestures which produced them. Furthermore, it is not necessary to adapt the articulatory pattern in /t/ because the necessary information will be there in the signal even without the adaptation. Conversely, for the fricative an articulatory adaptation is necessary because otherwise the information "alveolar constriction" cannot be recovered from the acoustic signal.

However, it is not possible to explain why speakers should vary the articulation by keeping the acoustic output constant as shown for /u/. If speakers aim at transmitting articulatory information there is no reason to vary exactly this parameter. Also, the adjustment of the jaw in the production of the alveolar fricative is difficult to explain. Thinking in articulatory terms, speakers should aim at transmitting the information "alveolar constriction". From this point of view it is not necessary to have a second obstacle at all. Speakers could therefore just ignore the missing second obstacle. Of course, one could argue, even if according to my knowledge nobody has done

so far, that another piece of articulatory information which is transmitted in /s/ should be "(upper) incisors as second obstacle". However, then it is not clear, why speakers should use the lower incisors exclusively in order to compensate for the non-availability of the upper ones and transmit the information that the upper incisors are used. Similarly to what was observed for the motor equivalent strategies in the production of /u/ one can see that speakers here vary the articulation but keep the acoustic output constant.

Still, there are a number of findings which cannot be explained by an adaptation towards acoustic phoneme targets. For example, speakers can adapt very well when no auditory feedback is there. The vowels are nearly perfect in this condition, but the fricatives as well can be classified in a reasonable way, even if perceptual quality estimations should show that they are not like the unperturbed productions. Thus, speakers must have an articulatory representation, even if it might only be one which has a motor function, which is used at least as long as no auditory feedback is there.

A comparison with the study by Jones and Munhall [2003] might be used to illustrate the point that articulatory representations are used when no auditory feedback is available. As described several times here, Jones & Munhall prolonged the teeth and investigated adaptation in /s/. They found that speakers did not adapt until auditory feedback was available. As shown here, when speakers are confronted with a palatal prosthesis, this is different. They adapt even when no auditory feedback is there. More concretely, during /s/ they lower and retract the tongue a little even in session E1/2. In contrast to the earlier study, for the speakers of the present study it can be assumed that they explore the characteristics of the perturbation (which the speakers in the earlier study could not easily do) by using tactile feedback and then estimate a new articulatory position. The retraction could be due to an articulatory representation in form of a linguo-palatal contact pattern which the speakers associate with the sound. The speakers might thus change the tongue shape in order to reach a certain contact pattern with the new palate shape. Another possibility would be that the articulatory representation is rather a certain constriction size. Speakers could then notice that the constriction is too narrow when the prosthesis is there, so that they lower and retract the tongue. While all these ideas are very hypothetical, a main point, for which there is evidence in the data, is, that speakers must use some kind of articulatory information while adapting without auditory feedback.

12.2.2 Motor aspects of adaptation

As stated in the introduction, articulatory representations could belong to the common currency, the information which is transmitted from the speaker to

the listener. Alternatively, they could have pure motor function and therefore exist in the speech production tasks only, but not in the perceptual primitives. It would therefore be useful to separate the two. Probably this separation cannot be carried out with certainty here, but still the data presented here might support one or the other argumentation.

The existence of motor equivalent strategies has often been interpreted as follows: There is considerable variability in the articulation but less in the acoustic output, so the perceptual information must be in the acoustic domain. Conversely, one could assume that if there is only little variability in the articulation then this should support articulatory representations. In fact, it often seems that the articulatory variability is quite low; in unperturbed speech speakers do not show very many instances of motor equivalent strategies. A question resulting from that could be, if the perceptual representations are in the acoustics, why do speakers not make more use of motor equivalent strategies? Why, for example, do so few speakers compensate via larynx lowering for less lip protrusion, why does nobody produce an /u/ with a velo-pharyngeal constriction and open lips, if it all results in the same acoustic output? The reason could simply be that the latter involves too much effort rather than that the articulatory information "labial and velar constriction" must be transmitted. Thus, speakers do not use all the strategies available because they select the one which involves the least effort which is then their articulatory representation of the sound. This assumption is supported by our results found for movement optimisation. Even if they are weak they still suggest that movement optimisation exists. The selection of a particular articulatory strategy could be the reason for the low articulatory variability which is often found in speech production.

Another argument for articulatory motor representations comes from the fact that trade-offs, which have been found for perception, are not used by the speakers in our study in order to adapt. Speakers, for example, do not shorten the lax vowel /ɪ/, and they do not produce longer /e/ and /o/ when the formant values deviate from the "ideal". This could again be because they have already decided on a way to produce the sound so that, even if such a production (with a different duration) would be perceptually acceptable, it just does not occur.

When speakers have selected a certain optimal strategy they might remember this strategy and it might serve for motor purposes. They might use it when no auditory feedback is there, but it is not transmitted to the listener. Thus, what was assumed in the *Auditory Enhancement Theory* for speech perception might have only restricted importance for speech production.

As stated above, an articulatory representation could potentially be a

contact pattern, a constriction size or an articulatory gesture. Evidence for one or the other seems to be in the data. For example, it was found that /i/ seems to improve less when auditory feedback becomes available than /e/. This could be because in the high vowel the use of a contact pattern is possible whereas in the intermediately high vowel it is not. Further support for a representation in terms of a contact pattern comes from the fact that no difference could be found between the two prosthesis types. When speakers have a contact pattern in mind they can realise it, no matter what the new palate shape is. Representations in terms of constriction sizes should also be easier to realise when there is lots of linguo-palatal contact. Gestural representations could be involved in the first stages of the adaptation when probably just the jaw is lowered. Afterwards, however, when the movements are reorganised, these gestural representations, if they exist, must change.

The idea that articulatory representations have motor function only could even go together with the idea that we can perceive motor plans via a mirror neuron system (see chapter 1). The information we receive from this system might not be linguistic, it might well be motor plans with only motor function.

12.2.3 Reparametrisation and reorganisation

As discussed in chapter 3 for the present experiment an initial reparametrisation followed by a reorganisation of the movement could be expected. A reparametrisation strategy would use the underlying characteristics of the habitual strategy, but would give different weights to several components of the strategy. Reparametrisation could be seen in the fricatives where the tongue was initially lowered and retracted without further changes such as an adaptation of the jaw position. Further evidence for a very long reparametrisation period could be found while looking at the development of the jerk values in /o/. Changes in jerk suggest a reorganisation of the movement. For two speakers, however, there was a very long phase where these values stayed constant. These speakers thus did not reorganise their movements, but they just reparametrised.

Reorganisation, on the other hand, means that speakers give up the initially used strategy and develop a new one which (possibly) results in a better acoustic output. Reorganisation occurs later, for example, when speakers raise the jaw in /s/ in session E1/3. Initially speakers lower the jaw because there is the prosthesis which reduces the space in the mouth and they want to compensate for that. This is a reparametrisation. Speakers give a new "height value" to the jaw so that the articulatory configuration as such stays the same. Only later do they change the articulatory configuration by raising

the jaw and thus putting it in a different relation to the tongue.

Reorganisation takes very long, and probably the adaptation time was still too short in order to adapt some of the fricatives. This is supported by the findings in Heydecke et al. [2004] who report that only speakers who had worn dental devices for several years managed to compensate completely for palatal coverage.

12.2.4 Speakers' aims and how they achieve them

In conclusion, speech production tasks contain both articulatory and acoustic components. The results presented here speak for the acoustic component as the essential component which is perceptually relevant, otherwise motor equivalent strategies would not have been found. This acoustic representation is not a fixed target, as can be seen by comparing acoustic measurements across sessions. What seems to be important is that the phonemes are distinct enough, in fact it was found that there is a certain tendency towards a maximisation of the space between the sounds of the German vowel system.

However, speakers at the same time dispose of an articulatory representation which seems to have a motor function. This articulatory representation seems to be extremely flexible, as can be seen by the compensatory abilities speakers show when no auditory feedback is available. The articulatory representation therefore cannot be something static as for example a gestural representation because even in this early stage speakers adapted via changing the tongue shape. In order to do this, speakers might have used articulatory representations such as for example tongue-palate contact patterns. For the production of sounds without linguo-palatal contact, however, they must dispose either of experience of speaking under perturbation or of a mechanism to estimate the necessary articulatory configuration in order to produce a certain acoustic output.

When speakers are first confronted with the perturbation they are thus able to produce utterances which can be classified according to acoustic parameters. When auditory feedback becomes available speakers change the more fine-grained acoustic properties of the sound, for example, they raise the jaw in order to produce the characteristic high-frequency noise in fricatives. Obviously, not all measureable acoustic parameters have the same importance. For the sibilants it seems to be more important to have this high-frequency noise than to have the original centre of gravity.

The adaptation process at first involves a reparametrisation which is followed by a reorganisation. During the reorganisation the articulatory representation is presumably replaced by a new one which represents an optimal strategy to reach the acoustic output.

Bibliography

W.A. Aasland, S. Baum, and D.H. McFarland. Electropalatographic, acoustic and perceptual data on adaptation to a palatal perturbation. *Journal of the Acoustical Society of America*, 119(4):2372–2380, 2006.

S. Baum and D.H. McFarland. The development of speech adaptation to an artificial palate. *Journal of the Acoustical Society of America*, 102(4): 2353–2359, 1997.

S. Baum and D.H. McFarland. Individual differences in speech adaptation to an artificial palate. *Journal of the Acoustical Society of America*, 107 (6):3572–3575, 2000.

S.R. Baum, D.H. McFarland, and M. Diab. Compensation to articulatory perturbation: Perceptual data. *Journal of the Acoustical Society of America*, 99(6):3791–3794, 1996.

S. Blumstein and K. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society America*, 66:1001–1017, 1979.

P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer, 1992–2004. URL `www.PRAAT.org`.

J. Bortz. *Statistik für Sozialwissenschaftler*. Springer, Berlin, Heidelberg, New York, 1999.

J. Brunner, P. Hoole, and P. Perrier. Articulatory optimisation in perturbed vowel articulation. In *Proceedings of the 26th ICPhS*, pages 497–500, 2007.

F.S. Cooper, P.C. Delattre, A.M. Liberman, J.M. Borst, and L.J. Gerstman. Some experiments on the perception of speech sounds. *Journal of the Acoustical Society of America*, 24:597–606, 1952.

R. Daniloff, G. Schuckers, and L. Feth. *The Physiology of Speech and Hearing: An introduction*. Prentice-Hall, Engelswood Cliffs NJ, 1980.

R. L. Diehl and K. R. Kluender. On the objects of speech perception. *Ecological Psychology*, 1:121–144, 1989.

R.L. Diehl and J. Kingston. Phonetic covariation as auditory enhancement: The case of the [+voice]/[-voice] distinction. In O. Engstrand and C. Kylander, editors, *Current Phonetic Research Paradigms: Implications for Speech Motor Control, PERILUS*, volume 14, pages 139–143. University of Stockholm, Stockholm, 1991.

R. M. W. Dixon. *The languages of Australia*. CUP, Cambridge, 1980.

V. Evers, H. Reeth, and A. Lahiri. Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics*, 26: 345–370, 1998.

L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti. Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, 15:399–402, 2002.

Gunnar Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

H. Fiokowski. *Sprecherzieherisches Elementarbuch*. Niemeyer, Tübingen, 2004.

J. Flege, S. Fletcher, and A. Homiedan. Compensating for a bite block in /s/ and /t/ production: Palatographic, acoustic, and perceptual data. *Journal of the Acoustical Society of America*, 83:212–228, 1988.

J.W. Folkins and G.N. Zimmermann. Lip and jaw interaction during speech: Responses to perturbatin of lower-lip movement prior to bilabial closure. *Journal of the Acoustical Society of America*, 71(5):1225–1233, 1982.

K. Forrest, G. Weismer, P. Milencovic, and R. N. Dougall. Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society America*, 84(1):115–123, 1988.

C. Fowler. Calling a mirage a mirage: direct perception of speech produced without a tongue. *Journal of Phonetics*, 18:529–541, 1990.

C. A. Fowler. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3):1730–1741, 1996.

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

C.A. Fowler and J.M. Brown. Intrinsic f0 differences in spoken and sung vowels and their perception by listeners. *Perception and Psychophysics*, 59 (5):729–738, 1997.

C.A. Fowler and D.J. Dekle. Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17:816–828, 1991.

C.A. Fowler, P. Rubin, R.E. Remez, and M.T. Turvey. Implications for speech production of a general theory of action. In B. Butterworth, editor, *Language Production*, pages 373–420. Academic Press, New York, 1980.

T. Gay, L.-J. Boë, and P. Perrier. Acoustic and perceptual effects of changes in vocal tract constrictions for vowels. *Journal of the Acoustical Society of America*, 92(3):1301–1309, 1992.

L.M. Goldstein and C. Fowler. Articulatory phonology: a phonology for public language use. In A.S. Meyer and N.O. Schiller, editors, *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Mouton de Gruyter, Berlin, 2003.

V.L. Gracco and J.H. Abbs. Dynamic control of the perioral system during speech: Kinematic analyses of autogenic and nonautogenic sensorimotor processes. *Journal of Neurophysiology*, 54(2):418–432, 1985.

F. Guenther. Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production. *Psychological Review*, 102(3): 594–621, 1995.

S.L. Hamlet and M. Stone. Compensatory vowel characteristics resulting from the presence of different types of experimental dental prostheses. *Journal of Phonetics*, 4:199–218, 1976.

S.L. Hamlet and M. Stone. Compensatory alveolar consonant production induced by wearing a dental prosthesis. *Journal of Phonetics*, 6:227–248, 1978.

K.S. Harris. Physiological aspects of articulatory behavior. In T.A. Sebeok, editor, *Current Trends in Linguistics*, volume 12, pages 2281–2302. Mouton, The Hague, 1974.

G. Heydecke, D.H. McFarland, J.S. Feine, and J.P. Lund. Speech with maxillary implant prostheses: Ratings of articulation. *Journal of Dental Research*, 83(3):236–240, 2004.

N. Hogan. An organising principle for a class of voluntary movements. *Journal of Neuroscience*, 4:2745–2754, 1984.

M. Honda and T. Kaburagi. Speech compensation to dynamical structural perturbation of the palate shape. In *Proc. of the 5th Speech Production Seminar and CREST Workshop on Models of Speech Production*, pages 21–24, Kloster Seeon, Bavaria, 2000.

M. Honda, A. Fujino, and T. Kaburagi. Compensatory responses of articulators to unexpected perturbation of the palate shape. *Journal of Phonetics*, 30:281–302, 2002.

P. Hoole. *Experimental studies of laryngeal articulation*. University of Munich, 2006. Habilitation.

P. Hoole. Theoretische und methodische Grundlagen der Artikulationsanalyse in der experimentellen Phonetik. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 34: 3–173, 1996a.

P. Hoole. Issues in the acquisition, processing, reduction and parameterization of articulographic data. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 34:158–173, 1996b.

P. Hoole. *mt_new*, 2007. URL `www.phonetik.uni-muenchen.de/~hoole`.

P. Hoole and C. Mooshammer. Articulatory analysis of the german vowel system. In P. Auer, P. Gilles, and H. Spiekermann, editors, *Silbenschnitt und Tonakzente*, pages 129–152. Niemeyer, Tübingen, 2002.

P. Hoole, A. Zierdt, and C. Geng. Beyond 2D in articulatory data acquisition and analysis. In *Proc. of the 15th ICPhS*, pages 265–268, 2003.

J.F. Houde and M.I. Jordan. Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, 45:295–310, 2002.

R. Jakobson, C. G. M. Fant, and M. Halle. *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, Mass.: MIT Press, 1961.

M. Jeannerod. The representing brain. Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17:187–245, 1994.

M. Jessen. *Phonetics and phonology of tense and lax obstruents in German*. Studies in Functional and Structural Linguistics 44. Benjamins, Philadelphia, 1999.

K. Johnson. The role of perceived speaker identity in f0 normalization of vowels. *Journal of the Acoustical Society of America*, 88:642–654, 1989.

J.A. Jones and K.G. Munhall. Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Journal of the Acoustical Society of America*, 113(1):532–543, 2003.

J.A. Jones and K.G. Munhall. Perceptual calibration of F0 production: evidence from feedback perturbation. *Journal of the Acoustical Society of America*, 108(3):1246–1251, 2000.

M.I. Jordan. Computational aspects of motor control and motor learning. In H. Heuer and S. Keele, editors, *Handbook of Perception and Action: Motor Skills*, pages 71–120. Academic Press, New York, 1996.

J.A.S. Kelso and B. Tuller. "Compensatory articulation" under conditions of reduced afferent information: A dynamic formulation. *Journal of Speech and Hearing Research*, 26:217–224, 1983.

C. Keysers, E. Kohler, M.A. Umiltà, L. Nanetti, L. Fogassi, and V. Gallese. Audiovisual mirror neurons and action recognition. *Experimental Brain Research*, 153(4):628–636, 2003.

J. Kingston and R. Diehl. Intermediate properties in the perception of distinctive feature values. In B. Connell and A. Arvantini, editors, *Papers in Laboratory Phonology IV: Phonology and Phonetic evidence*, pages 7–27. Cambridge: CUP, 1995.

J. C. Kingston and R. Diehl. Phonetic knowledge. *Language*, 70:419–454, 1994.

K.R. Kluender. Effects of first formant onset properties on voicing judgments result from processes not specific to humans. *Journal of the Acoustical Society of America*, 90:83–96, 1991.

K.R. Kluender, R.L. Diehl, and R.R. Killeen. Japanese quail can learn phonetic categories. *Science*, 237:1195–1197, 1987.

R. Laboissière, D.J. Ostry, and A.G. Feldman. The control of multi-muscle systems: Human jaw and hyoid movements. *Biological Cybernetics*, 74: 373–384, 1996.

Alvin M. Liberman and Ignatius G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21:1–36, 1985.

A.M. Liberman. Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, and N. Thorsen, editors, *Proc. of the IXth ICPhS*, volume 2, pages 468–473, Copenhagen, 1979. University of Copenhagen.

A.M. Liberman, F.S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological Review*, 74:431–461, 1967.

J. Liljencrants and B.E.F. Lindblom. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, 48:839–862, 1972.

B. Lindblom. Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99(3):1683–1692, 1996.

B. Lindblom. Phonetic universals in vowel systems. In John J. Ohala and Jeri J. Jaeger, editors, *Experimental Phonology*, pages 13–44. Academic Press, Orlando, 1986.

B. Lindblom. Phonetic invariance and the adaptive nature of speech. In Ben A.G. Elsendoom and H. Bouma, editors, *Working Models of Human Perception*, pages 139–173. Academic Press, London, 1988.

B. Lindblom. Explaining phonetic variation: A sketch of the H&H theory. In William J. Hardcastle and Alain Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers, Dordrecht, 1990.

B. Lindblom, J. Lubker, and T. Gay. Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, 7:147–161, 1979.

A. Löfqvist, T. Baer, N.S. McGarr, and R.S. Story. The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America*, 85:1314–1321, 1989.

P.F. MacNeilage and J.L. deClerk. On the motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America*, 45: 1217–1233, 1969.

D.H. McFarland and S.R. Baum. Incomplete compensation to articulatory perturbation. *Journal of the Acoustical Society of America*, 97:1865–1873, 1995.

D.H. McFarland, S. Baum, and C. Chabot. Speech compensation to structural modifications of the oral cavity. *Journal of the Acoustical Society of America*, 100(2):1093–1104, 1996.

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

J. D. Miller. Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85(5):2114–2134, 1989.

C. Mooshammer, P. Hoole, and A. Geumann. Interarticulator cohesion within coronal consonant production. *Journal of the Acoustical Society of America*, 120(2):1028–1039, 2006.

E.C.E. Morrish. The direct-realist theory of speech perception: counter-evidence from the analysis of the speech of a glossectomee. *Journal of Phonetics*, 18:519–527, 1990.

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

W. L. Nelson. Physical principles for economies of skilled movements. *Biological Cybernetics*, 46:135–147, 1983.

J.J. Ohala. Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3):1718–1725, 1996.

J. Perkell, F. Guenther, H. Lane, M. Matthies, P. Perrier, J. Vick, R. Wilhelms-Tricarico, and M. Zandipour. A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, 28(3):233–272, 2000.

J.S. Perkell, M.L. Matthies, M.A. Svirsky, and M.I. Jordan. Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot 'motor equivalence' study. *Journal of the Acoustical Society of America*, 93(5):2948 – 2961, 1993.

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

P. Perrier. Control and representations in speech production. *ZAS Papers in Linguistics*, 40:109–132, July 2005.

G.E. Peterson and H.L. Barney. Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1952.

C. Pörschmann. Influences of bone conduction and air conduction on the sound of one's own voice. *Acoustica - acta acoustica*, 86(6):1038–1045, 2000.

D. W. Purcell, H. Kunov, and W. Cleghorn. Estimating bone conduction transfer functions using otoacoustic emissions. *Journal of the Acoustical Society of America*, 114(2):907–918, 2003.

D.W. Purcell and K.G. Munhall. Compensation following real-time manipulation of formants in isolated vowels. *Journal of the Acoustical Society of America*, 119(4):2288–2297, 2006a.

D.W. Purcell and K.G. Munhall. Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, 120(2):966–977, 2006b.

N. Reinholt-Peterson. Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations. *Phonetica*, 43:31–42, 1986.

G. Rizzolatti and M.A. Arbib. The representing brain. Neural correlates of motor intention and imagery. *Trends in Neurosciences*, 21:188–194, 1998.

G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

E.L. Saltzman and K.G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382, 1989.

C. Savariaux, P. Perrier, and J.-P. Orliaguet. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production. *Journal of the Acoustical Socitey of America*, 98:2428–2442, 1995.

C. Savariaux, P. Perrier, J.-P. Orliaguet, and J.-L. Schwartz. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube II. Perceptual analysis. *Journal of the Acoustical Society of America*, 106: 381–393, 1999.

C. Shadle. The acoustics of fricative consonants. Technical Report 506, Research Laboratory of Electronics, MIT Cambridge, MA., 1985.

C. Shadle. The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics*, 19:409–424, 1991.

D.M. Shiller, D.J. Ostry, and P.L. Gribble. Effects of gravitational load on jaw movements in speech. *Journal of Neuroscience*, 19:9073–9080, 1999.

D.M. Shiller, D.J. Ostry, P.L. Gribble, and R. Laboissière. Compensation for the effects of head acceleration on jaw movements in speech. *Journal of Neuroscience*, 15:6447–6456, 2001.

D.M. Shiller, R. Laboissière, and D.J. Ostry. Relationship between jaw stiffness and kinematic variability in speech. *Journal of Neurophysiology*, 88: 2329–2340, 2002.

K. Silverman. *The structure and processing of fundamental frequency contours*. Cambridge University, 1987. doctoral dissertation.

K. Stevens. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. Davis and Peter Denes, editors, *Human Communication: A Unified View*, pages 51–66. McGraw-Hill, New York, 1972.

K. Stevens. Models of phonetic recognition II: A feature-based model of speech recognition. In P. Mermelstein, editor, *Proc. of the Montréal Satellite Symposium on Speech Recognition*, pages 67–68, 1986.

K. Stevens. Phonetic features and lexical access. In *The Second Symposium on Advanced Man-Machine Interface Through Spoken Language*, pages 10–23, Japan, 1988.

K. Stevens and S. Blumstein. The search for invariant acoustic correlates of phonetic features. In P. Eimas and J. Miller, editors, *Perspectives on the Study of Speech*, pages 1–38. Erlbaum, Hillsdale, 1981.

K. Stevens and S. Blumstein. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64:1358–1368, 1978.

K. Stevens, S.Y. Manuel, S. Shattuck-Hufnagel, and S. Liu. Implementation of a model for lexical access based on features. In J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, and G.E. Weibe, editors, *Proceedings ICSLP-92*, volume 1, pages 499–522, 1992.

W. Strange, R.R. Verbrugge, D.P. Shankweiler, and T.R. Edman. Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 60:213–224, 1976.

M. Tabain and A. Butcher. Stop consonants in Yanyuwa and Yindjibarndi: A locus equation perspective. *Journal of Phonetics*, 27:333–357, 1999.

M. Tiede, S. Masaki, and E. Vatikiotis-Bateson. Contrasts in speech articulation observed in sitting and supine position. In *Proc. 5th Speech Production Seminar*, pages 25–28, 2000.

S. Tremblay, D.M. Shiller, and D.J. Ostry. Somatosensory basis of speech production. *Nature*, 423:866–869, 2003.

B. Tuller and J. Kelso. The timing of articulatory gestures: Evidence for relational invariants. *Journal of the Acoustical Society of America*, 76(4): 1030–1036, 1984.

B. Tuller, J. Kelso, and K. Harris. Interarticulator phasing as an index of temporal regularity in speech. *Journal of Experimental Psychology*, 8(3): 460–472, 1982.

B. Tuller, J. Kelso, and K. Harris. Converging evidence for the role of relative timing in speech. *Journal of Experimental Psychology*, 9(5):829–833, 1983.

M.A. Umiltà, E. Kohler, V. Gallese, L. Fogassi, and L. Fadiga. "I know what you are doing": a neurophysiological study. *Neuron*, 32:91–101, 2001.

Y. Uno, M. Kawato, and R. Suzuki. Formation and control of optimal trajectory in human multijoint arm movement - minimum torque-change model. *Biological Cybernetics*, 61:89–101, 1989.

T. Vennemann. Skizze der deutschen Wortprosodie. *Zeitschrift für Sprachwissenschaft*, 10:86–111, 1991.

R.S. Waldstein. Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. *Journal of the Acoustical Society of America*, 88:2099–2114, 1990.

K.E. Watkins, A.P. Strafella, and T. Paus. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41: 989–994, 2003.

C. Watson and J. Harrington. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106(1):458–468, 1999.

J. Westbury, M. Hashi, and M. Lindstrom. Differences among speakers in lingual articulation of American English /r/. *Speech Communication*, 26: 203–226, 1998.

J.R. Westbury, M.J. Lindstrom, and M.D. McClean. Tongues and lips without jaws: A comparison of methods for decoupling speech movements. *Journal of Speech, Language, and Hearing Research*, 45:651–662, 2002.

X. Zhou, C. Espy-Wilson, M. Tiede, and S. Boyce. Acoustic cues of "retroflex" and "bunched" American English rhotic sound. *Journal of the Acoustical Society of America*, 121(5):3168, 2007.

W. Ziegler and P. Hoole. A combined acoustic and perceptual analysis of the tense-lax opposition in aphasic vowel production. *Aphasiology*, 3(5): 449–463, 1989.

A. Zierdt, P Hoole, and H.G. Tillmann. Development of a system for three-dimensional fleshpoint measurement of speech movements. In *Proc. of the 14th ICPhS*, pages 73–75, 1999.

A. Zierdt, P Hoole, M. Honda, T. Kaburagi, and H.G. Tillmann. Extracting tongues from moving heads. In *Proc. of the 5th Speech Production Seminar*, pages 313–316, 2000.

V.W. Zue. The use of speech knowledge in automatic speech recognition. In *Proceedings IEEE*, volume 73, pages 1602–1615, 1985.

Appendix A

Statistics

Table A.1: Influence of the vowel identity and the session on the formant values. Results of the repeated measures ANOVA calculated for data split by speaker with vowel and session as factors and the formants as dependent variable. Column 1: speaker, column 2: effect, columns 3-5: F values with Greenhouse-Geisser corrected degrees of freedom and significance levels. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. For further details see section 5.1.

speaker	effect	F1	F2	F3
AP	vowel	F(5, 929)=169.984 ***	F(5, 916)= 7615.444 ***	F(5, 919)= 512.256 ***
	session	F(7, 929)=24.035 ***	F(7, 916)= 14.694 ***	F(7, 919)= 99.892 ***
	vowel*session	F(35, 929)= 3.238 ***	F(35, 916)= 12.989 ***	F(35, 919)= 15.901 ***
BP	vowel	F(5, 951)=633.586 ***	F(5, 948)= 15760.681 ***	F(5, 936)= 872.717 ***
	session	F(7, 951)=39.147 ***	F(7, 948)= 30.253 ***	F(7, 936)= 59.456 ***
	vowel*session	F(35, 951)= 4.150 ***	F(35, 948)= 36.651 ***	F(35, 936)= 23.4 ***
DS	vowel	F(5, 947)=809.788 ***	F(5, 937)= 10321.977 ***	F(5, 927)= 297.033 ***
	session	F(7, 947)=7.621 ***	F(7, 937)= 23.679 ***	F(7, 927)= 68.437 ***
	vowel*session	F(35, 947)= 12.118 ***	F(35, 937)= 5.496 ***	F(35, 927)= 9.236 ***
KD	vowel	F(5, 815)=1331.020 ***	F(5, 815)= 27407.933 ***	F(5, 810)= 3479.429 ***
	session	F(6, 815)=12.970 ***	F(6, 815)= 15.919 ***	F(6, 810)= 111.082 ***
	vowel*session	F(30, 815)= 5.701 ***	F(30, 815)= 5.303 ***	F(30, 810)= 35.443 ***
OP	vowel	F(5, 785)=627.460 ***	F(5, 784)= 27835.816 ***	F(5, 778)= 1111.321 ***
	session	F(6, 785)=7.041 ***	F(6, 784)= 17.711 ***	F(6, 778)= 86.32 ***
	vowel*session	F(30, 785)= 2.037 ***	F(30, 784)= 9.553 ***	F(30, 778)= 13.316 ***
SK	vowel	F(5, 1037)=282.725 ***	F(5, 1031)= 14975.616 ***	F(5, 1038)= 2197.105 ***
	session	F(7, 1037)=8.755 ***	F(7, 1031)= 38.554 ***	F(7, 1038)= 107.657 ***
	vowel*session	F(35, 1037)= 3.689 ***	F(35, 1031)= 6.996 ***	F(35, 1038)= 16.554 ***
TP	vowel	F(5, 950)=346.168 ***	F(5, 931)= 19548.471 ***	F(5, 943)= 2722.843 ***
	session	F(7, 950)=26.285 ***	F(7, 931)= 42.045 ***	F(7, 943)= 39.04 ***
	vowel*session	F(35, 950)= 8.334 ***	F(35, 931)= 18.671 ***	F(7, 35)= 11.254 ***

Table A.2: Overlap in percent between F-values of different vowels for AP and BP. Column 1: speaker, column 2: vowel with respect to which overlap is calculated, column 3: formant, columns 4 and 5: boundaries of formant regions which are defined by means ± 2 std. dev. of the measured F-values over sessions. Columns 6 to 11: overlap in percentage of a 2nd vowel with respect to the one in column 2. For further details see section 5.1.

			boundaries		overlap					
speaker	vowel	formant	lower	upper	e	ɪ	i	o	u	y
AP	e	F1	278	524	100	**87**	**49**	87	85	48
	e	F2	2332	2835	100	**57**	**95**	0	0	0
	e	F3	2641	3907	100	**89**	**99**	58	44	20
	ɪ	F1	306	519	**100**	100	**43**	98	85	42
	ɪ	F2	2160	2617	**62**	100	**64**	0	0	0
	ɪ	F3	2701	3829	**100**	100	**100**	60	44	18
	i	F1	224	398	**69**	**53**	100	51	100	95
	i	F2	2323	2811	**98**	**60**	100	0	0	0
	i	F3	2655	3992	**94**	**84**	100	54	40	18
	o	F1	310	546	91	89	37	100	**75**	36
	o	F2	562	1159	0	0	0	100	**85**	0
	o	F3	2440	3381	79	72	77	100	**80**	49
	u	F1	200	488	73	63	60	**62**	100	58
	u	F2	650	1299	0	0	0	**78**	100	0
	u	F3	2226	3192	57	51	56	**78**	100	70
	y	F1	229	395	70	54	100	51	100	100
	y	F2	1540	2135	0	0	0	0	0	100
	y	F3	2210	2900	38	29	36	67	98	100
BP	e	F1	346	448	100	**25**	**14**	100	31	15
	e	F2	1716	2053	100	**64**	**58**	0	0	33
	e	F3	2166	2762	100	**81**	**84**	77	50	11
	ɪ	F1	290	371	**31**	100	**86**	31	100	88
	ɪ	F2	1838	2120	**76**	100	**93**	0	0	0
	ɪ	F3	2279	3092	**59**	100	**100**	51	29	0
	i	F1	276	360	**17**	**83**	100	17	100	100
	i	F2	1857	2213	**55**	**74**	100	0	0	0
	i	F3	2262	3221	**52**	**85**	100	45	27	0
	o	F1	346	475	79	19	11	100	**25**	12
	o	F2	586	913	0	0	0	100	**100**	0
	o	F3	2235	2695	100	90	94	100	**61**	0
	u	F1	255	378	26	66	68	**26**	100	75
	u	F2	584	1103	0	0	0	**63**	100	0
	u	F3	2218	2517	100	80	85	**94**	100	4
	y	F1	269	361	**16**	77	91	16	100	100
	y	F2	1477	1828	**32**	0	0	0	0	100
	y	F3	1858	2230	**17**	0	0	0	3	100

Table A.3: As table A.2, but for speaker DS.

			boundaries		overlap					
speaker	vowel	formant	lower	upper	e	ɪ	i	o	u	y
DS	e	F1	339	426	100	**100**	**13**	89	9	**18**
	e	F2	1732	1943	100	**47**	**99**	0	0	**29**
	e	F3	2419	3113	100	**92**	**100**	44	75	**74**
	ɪ	F1	330	431	**86**	100	**20**	77	17	**25**
	ɪ	F2	1593	1832	**42**	100	**41**	0	0	**84**
	ɪ	F3	2475	3129	**98**	100	**100**	47	71	**70**
	i	F1	270	350	**14**	**25**	100	1	89	**84**
	i	F2	1734	2002	**78**	**37**	100	0	0	**22**
	i	F3	2392	3202	**86**	**81**	100	38	68	**67**
	o	F1	349	427	99	100	1	100	0	8
	o	F2	705	953	0	0	0	100	77	0
	o	F3	2640	2945	100	100	100	100	99	95
	u	F1	276	347	11	24	100	0	100	90
	u	F2	763	1136	0	0	0	51	100	0
	u	F3	2021	2942	57	51	60	33	100	99
	y	F1	283	355	**22**	**35**	**93**	8	89	100
	y	F2	1576	1793	**28**	**92**	**27**	0	0	100
	y	F3	1722	2931	**42**	**38**	**45**	24	75	100

Table A.4: As table A.2, but for speakers KD and OP.

			boundaries		overlap					
speaker	vowel	formant	lower	upper	e	ɪ	i	o	u	y
KD	e	F1	386	592	100	**87**	0	77	0	0
	e	F2	2407	2620	100	**69**	83	0	0	0
	e	F3	2905	3440	100	**96**	69	0	0	0
	ɪ	F1	373	565	**93**	100	0	82	0	0
	ɪ	F2	2242	2555	**47**	100	36	0	0	0
	ɪ	F3	2924	3536	**84**	100	76	0	0	0
	i	F1	279	323	0	0	100	0	70	82
	i	F2	2443	2715	65	41	100	0	0	0
	i	F3	3069	4079	37	46	100	0	0	0
	o	F1	407	566	100	99	0	100	0	0
	o	F2	738	934	0	0	0	100	100	0
	o	F3	2459	2892	0	0	0	100	86	37
	u	F1	292	343	0	0	61	0	100	73
	u	F2	698	999	0	0	0	65	100	0
	u	F3	2503	2874	0	0	0	100	100	32
	y	F1	287	329	0	0	86	0	88	100
	y	F2	1713	1952	0	0	0	0	0	100
	y	F3	2267	2620	0	0	0	46	33	100
OP	e	F1	403	486	100	**86**	0	92	0	0
	e	F2	2556	2874	100	**76**	76	0	0	0
	e	F3	3069	3451	100	**100**	94	30	38	0
	ɪ	F1	306	474	**42**	100	**30**	38	48	54
	ɪ	F2	2448	2799	**69**	100	**48**	0	0	0
	ɪ	F3	2973	3454	**79**	100	**75**	43	51	0
	i	F1	201	357	0	**33**	100	0	68	73
	i	F2	2631	2908	88	**61**	100	0	0	0
	i	F3	3093	3725	57	**57**	100	14	19	0
	o	F1	410	486	100	84	0	100	0	0
	o	F2	627	935	0	0	0	100	64	0
	o	F3	2637	3182	21	38	16	100	100	43
	u	F1	251	387	0	60	78	0	100	100
	u	F2	515	825	0	0	0	64	100	0
	u	F3	2431	3216	19	31	16	69	100	38
	y	F1	243	396	0	59	75	0	89	100
	y	F2	1555	1908	0	0	0	0	0	100
	y	F3	2575	2871	0	0	0	79	100	100

Table A.5: As table A.2, but for speakers SK and TP.

			boundaries		overlap					
speaker	vowel	formant	lower	upper	e	ɪ	i	o	u	y
SK	e	F1	332	465	100	**93**	**20**	100	95	62
	e	F2	2260	2586	100	**89**	**77**	0	0	0
	e	F3	2693	3396	100	**85**	**70**	34	25	0
	ɪ	F1	325	456	**95**	100	**26**	100	100	69
	ɪ	F2	2213	2551	**86**	100	**64**	0	0	0
	ɪ	F3	2801	3454	**91**	100	**84**	20	10	0
	i	F1	189	359	**16**	**20**	100	24	59	79
	i	F2	2335	2675	**74**	**64**	100	0	0	0
	i	F3	2907	3883	**50**	**56**	100	2	0	0
	o	F1	318	478	83	82	26	100	**88**	61
	o	F2	626	988	0	0	0	100	**100**	0
	o	F3	2341	2931	40	22	4	100	**89**	25
	u	F1	259	459	64	66	50	**71**	100	78
	u	F2	491	1071	0	0	0	**62**	100	0
	u	F3	2288	2866	30	11	0	**91**	100	35
	y	F1	225	415	44	47	71	51	82	100
	y	F2	1643	2115	0	0	0	0	0	100
	y	F3	2152	2488	0	0	0	44	60	100
TP	e	F1	269	373	100	**90**	**67**	70	48	64
	e	F2	1822	1995	100	**84**	**94**	0	0	0
	e	F3	2541	2926	100	**100**	**84**	38	0	0
	ɪ	F1	279	373	**100**	100	**64**	78	43	61
	ɪ	F2	1779	1968	**77**	100	**100**	0	0	0
	ɪ	F3	2449	3031	**66**	100	**73**	41	0	0
	i	F1	190	339	**47**	**40**	100	26	68	65
	i	F2	1751	1984	**70**	**81**	100	0	0	0
	i	F3	2604	3082	**67**	**89**	100	18	0	0
	o	F1	300	386	85	85	45	100	**22**	42
	o	F2	501	716	0	0	0	100	**100**	0
	o	F3	2282	2688	36	59	21	100	**27**	0
	u	F1	218	319	50	40	100	**19**	100	79
	u	F2	378	969	0	0	0	**36**	100	0
	u	F3	1887	2392	0	0	0	**22**	100	42
	y	F1	239	336	69	59	100	37	82	100
	y	F2	1466	1724	0	0	0	0	0	100
	y	F3	1802	2098	0	0	0	0	71	100

Table A.6: Influence of the vowel identity on the duration. Results of repeated measures ANOVA for data split by speaker. Column 1: speaker, column 2: effect, column 3: F-value, Greenhouse-Geisser corrected degrees of freedom and significance level. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. For further details see section 5.3

speaker	effect	F
TP	vowel	F(5, 935)= 184.106***
	session	F(7, 935)= 135.02***
	vowel*session	F(35, 935) 2.703***
KD	vowel	F(5, 809)= 408.784***
	session	F(6, 809)= 26.312***
	vowel*session	F(30, 809)= 3.244***
OP	vowel	F(5, 781)= 444.35***
	session	F(6, 781)= 19.376***
	vowel*session	F(30, 781)= 8.457***
SK	vowel	F(5, 1034)= 133.521***
	session	F(7, 1034)= 51.13***
	vowel*session	F(35, 1034)=6.142***
BP	vowel	F(5, 942)= 21.561***
	session	F(7, 942)= 3.248**
	vowel*session	F(35, 942)= 1.187 n.s.
AP	vowel	F(5, 924)= 203.364***
	session	F(7, 924)= 36.706***
	vowel*session	F(35, 924)= 1.923**
DS	vowel	F(5, 944)= 135.741***
	session	F(7, 944)= 15.898***
	vowel*session	F(35, 944)= 2.765***

Table A.7: Results of repeated measures ANOVA for the influence of the session on the articulatory and acoustic distance. Column 1: speaker, column 2: F-value, Greenhouse-Geisser corrected degrees of freedom and significance of influence on the articulatory (upper half) and acoustic (lower half) distance between /i/ and /u/, column 3: influence on the distance between /e/ and /o/. For further details see section 6.2.

speaker	art. /i-u/	art. /e-o/
TP	F(3.552, 19)= 144.854 ***	F(3.058, 19)=118.191 ***
SK	F(3.908, 19)=99.088 ***	F(3.611, 19)=78.155 ***
BP	F(3.640, 19)=139.344 ***	F(3.606, 19)=169.052 ***
AP	F(3.766, 19)=86.199 ***	F(2.951, 19)=126.921 ***
DS	F(3.547, 19)=2.882 ***	F(3.181, 19)=8.745 ***
KD	F(3.006, 19)=37.871 ***	F(3.479, 19)=60.380 ***
OP	F(3.055, 19)=64.082 ***	F(2.847, 19)=63.078 ***
speaker	ac. /i-u/	ac. /e-o/
TP	F(3.703, 19)=29.244 ***	F(4.009, 19)=7.628 ***
SK	F(3.338, 19)=6.443 ***	F(4.166, 19)=26.617 ***
BP	F(3.275, 19)=47.074 ***	F(4.119, 19)=46.715 ***
AP	F(3.437, 19)=31.546 ***	F(3.841, 19)=24.319 ***
DS	F(3.427, 19)=5.797 **	F(4.532, 19)=7.588 ***
KD	F(3.421, 19)=4.501 **	F(3.224, 19)=5.557 **
OP	F(3.590, 19)=25.135 ***	F(2.928, 19)=11.897 ***

Table A.8: Results of the repeated measures ANOVA for the influence of the condition (session) on the COG with Greenhouse-Geisser corrected degrees of freedom. Significance levels: *p<0.05, **p<0.01, ***p<0.001. Bold: The post-hoc tests resulted in a significant difference between all three conditions. Italics: The difference between the unperturbed (E1/1) and the perturbed session without feedback masking (E1/3) was significant. There is one exception: Speaker SK, /x/: Here all the differences were significant. However, the COG was higher in session E1/2 than in the other sessions. For further details see section 9.1

fricative	speaker	F
/s/	TP	F(1.7; 19)=1.443 n.s.
/s/	SK	*F(1.7; 19)=18.612****
/s/	BP	*F(1.9; 20)=15.974****
/s/	AP	F(1.9; 19)=7.223**
/s/	DS	**F(1.9; 20)=75.342*****
/ʃ/	TP	*F(1.8; 19)=10.932****
/ʃ/	SK	**F(1.7; 20)=34.610*****
/ʃ/	BP	F(1.8; 19)=3.594*
/ʃ/	AP	*F(1.8; 19)=48.499****
/ʃ/	DS	*F(1.8; 19)=37.139****
/ç/	TP	F(1.8; 20)=8.275**
/ç/	SK	F(1.9; 20)=6.829**
/ç/	BP	F(1.8; 19)=4.101*
/ç/	AP	**F(1.6; 19)=100.150*****
/ç/	DS	F(2.0; 19)=14.699***
/x/	TP	F(1.9; 19)=6.423**
/x/	SK	F(1.9; 20)=83.740***
/x/	BP	*F(1.8; 19)=8.324***
/x/	AP	**F(1.7; 19)=23.498*****
/x/	DS	*F(1.7; 19)=8.157***

Table A.9: Results of the repeated measures ANOVA for the influence of the condition (session) on the mean horizontal target position of the tongue with Greenhouse-Geisser corrected degrees of freedom. Significance levels: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$. Bold: The post-hoc tests resulted in a significant difference between all three conditions. Italics: The difference between the unperturbed (E1/1) and the perturbed session without feedback masking (E1/3) was significant. Underlined: The difference between E1/2 and E1/3 is significant. For further details see section 9.2

fricative	speaker	F
/s/	TP	*F(1.336, 19)=2.146 n.s.*
/s/	SK	F(1.507, 19)=2.174 n.s.
/s/	BP	F(1.848, 19)=142.085***
/s/	AP	**F(1.650, 19)=179.810*** **
/s/	DS	**F(1.706, 19)=209.274*** **
/ʃ/	TP	*F(1.111, 19)=44.054*** *
/ʃ/	SK	*F(1.858, 19)=25.822*** *
/ʃ/	BP	*F(1.865, 19)=37.087*** *
/ʃ/	AP	*F(1.959, 19)=48.184*** *
/ʃ/	DS	**F(1.804, 19)=38.301*** **
/ç/	TP	*F(1.708, 19)=10.232*** *
/ç/	SK	*F(1.974, 19)=88.001*** *
/ç/	BP	F(1.468, 19)=194.640***
/ç/	AP	*F(1.979, 19)=1619.244*** *
/ç/	DS	**F(1.591, 19)=300.856*** **
/x/	TP	*F(1.470, 19)=18.990*** *
/x/	SK	*F(1.422, 19)=11.740*** *
/x/	BP	F(1.555, 19)=147.607***
/x/	AP	*F(1.876, 19)=134.526*** *
/x/	DS	**F(1.979, 19)=15.819*** **

Table A.10: Cases in which the pattern *increase only* was found (vowels). Column 1: speaker, column 2: sound, column 3: sensor, columns 4-8: mean jerk values in different sessions, column 9: session with maximal jerk value, column 10: p of post-hoc test for the comparison session E1/1 - session with maximal value, column 11: p of post-hoc test for the comparison session with maximal value - session E3/1. For further details see section 11.2.

speaker	sound	sensor	E1/1	E1/2	E1/3	E2/1	E3/1	max	pI	pII
BP	/o/	ttip	1720(889)	1343(804)	1781(1336)	3372(1277)	2512(1107)	E2/1	0.001	0.286
OP	/o/	ttip	3262(1665)		3238(1584)	4019(1462)	5658(1944)	E3/1	0.003	
AP	/o/	jaw	749(557)	265(169)	493(285)	2158(1200)	1834(1295)	E2/1	0.001	0.996
BP	/o/	tdor	553(272)	504(292)	802(595)	3031(1192)	2046(964)	E2/1	0.000	0.075
KD	/o/	jaw	99(65)		325(180)	428(180)	549(264)	E3/1	0.000	
KD	/i/	ttip	109(92)		300(234)	339(215)	220(94)	E2/1	0.001	0.177
TP	/i/	tdor	168(96)	216(129)	124(68)	362(201)	233(224)	E2/1	0.006	0.494
DS	/i/	tdor	170(128)	147(88)	309(234)	338(205)	256(125)	E2/1	0.050	0.763
KD	/i/	tdor	154(218)		240(111)	198(117)	368(243)	E3/1	0.050	
TP	/i/	jaw	126(70)	268(152)	164(119)	129(61)	213(119)	E1/2	0.008	0.907
DS	/i/	jaw	90(61)	96(55)	162(130)	187(114)	125(68)	E2/1	0.028	0.400
KD	/i/	jaw	76(44)		79(41)	169(104)	95(68)	E2/1	0.007	0.072

Table A.11: Cases in which the pattern *increase only* was found (consonants). Column 1: speaker, column 2: sound, column 3: sensor, columns 4-8: mean jerk values in different sessions, column 9: session with maximal jerk value, column 10: p of post-hoc test for the comparison session E1/1 - session with maximal value, column 11: p of post-hoc test for the comparison session with maximal value - session E3/1. For further details see section 11.2.

speaker	sound	sensor	E1/1	E1/2	E1/3	E2/1	E3/1	max	pI	pII
DS	/z/	ttip	576(351)	835(426)	899(614)	1169(556)	986(370)	E2/1	0.003	0.936
OP	/z/	ttip	1390(579)		1831(953)	3032(1402)	2325(1743)	E1/3	0.000	0.663
SK	/z/	ttip	495(185)	1279(1061)	662(340)	1036(583)	908(337)	E1/2	0.039	0.808
DS	/z/	tdor	498(284)	556(344)	584(391)	974(423)	1008(516)	E3/1	0.006	
KD	/z/	tdor	633(299)		871(354)	1514(447)	1269(398)	E2/1	0.000	0.394
DS	/z/	jaw	254(123)	345(233)	291(173)	774(495)	657(353)	E2/1	0.002	0.993
KD	/z/	jaw	285(124)		496(217)	623(179)	609(205)	E2/1	0.000	1.000
TP	/z/	jaw	146(86)	297(146)	371(204)	255(170)	388(219)	E3/1	0.002	
SK	/t/	ttip	3202(1127)	7977(5347)	5500(2096)	4018(2235)	5227(2725)	E1/2	0.008	0.400
OP	/t/	ttip	3495(1227)		7797(2141)	6639(1915)	6536(1866)	E1/3	0.000	0.365
KD	/t/	tdor	1683(683)		2434(858)	4761(1367)	4318(950)	E2/1	0.000	0.823
OP	/t/	tdor	936(430)		2825(1284)	2332(957)	2016(702)	E1/3	0.000	0.185
TP	/t/	jaw	452(234)	1198(931)	985(511)	803(285)	746(318)	E1/2	0.022	0.409
KD	/t/	jaw	493(200)		1043(389)	1369(369)	1552(398)	E3/1	0.000	

Table A.12: Cases in which the pattern *increase-decrease* was found. Column 1: speaker, column 2: sound, column 3: sensor, columns 4-8: mean jerk values in different sessions, column 9: session with maximal jerk value, column 10: p of post-hoc test for the comparison session E1/1 - session with maximal value, column 11: p of post-hoc test for the comparison session with maximal value - session E3/1. For further details see section 11.2.

speaker	sound	sensor	E1/1	E1/2	E1/3	E2/1	E3/1	max	pI	pII
DS	/o/	ttip	2353(1436)	5973(4830)	7729(4749)	3613(1337)	1302(497)	E1/3	0.001	0.000
KD	/o/	ttip	1423(628)		3108(913)	4121(1026)	3077(1059)	E2/1	0.000	0.024
SK	/o/	ttip	1206(423)	3406(1458)	3379(1439)	927(507)	485(164)	E1/2	0.000	0.000
AP	/o/	ttip	1890(1105)	1129(561)	2610(1796)	7205(5676)	2629(1485)	E2/1	0.005	0.021
TP	/o/	tdor	4297(2603)	5052(2081)	5978(3805)	8936(3795)	4226(2210)	E2/1	0.001	0.000
DS	/o/	jaw	177(76)	220(141)	199(155)	549(254)	311(149)	E2/1	0.000	0.011
SK	/o/	jaw	140(104)	291(183)	159(132)	95(61)	110(89)	E1/2	0.031	0.005
TP	/o/	jaw	343(212)	1349(829)	624(467)	990(721)	482(332)	E1/2	0.000	0.002
AP	/i/	jaw	197(106)	152(110)	275(172)	837(383)	383(328)	E2/1	0.000	0.003
AP	/t/	ttip	28817(10245)	41350(17074)	66004(22162)	95057(35395)	60496(26724)	E2/1	0.000	0.015
DS	/t/	ttip	2943(2598)	5549(3040)	5762(1888)	5397(2487)	1766(735)	E1/3	0.004	0.000
KD	/t/	ttip	11697(3129)		3429(1297)	8268(1375)	6524(1430)	E2/1	0.001	0.002
TP	/t/	ttip	14340(5673)	19929(4295)	23513(6702)	12690(3649)	12855(6817)	E1/3	0.000	0.000
AP	/t/	tdor	544(302)	816(586)	843(410)	1383(593)	628(256)	E2/1	0.000	0.000
DS	/t/	tdor	646(360)	1093(532)	1844(1188)	2756(1292)	1666(797)	E2/1	0.000	0.030
SK	/t/	tdor	756(301)	1452(849)	1362(826)	916(515)	769(460)	E1/2	0.021	0.034
AP	/t/	jaw	334(174)	304(112)	588(249)	1785(1388)	606(374)	E2/1	0.002	0.014
DS	/t/	jaw	176(104)	248(142)	328(214)	772(376)	389(214)	E2/1	0.000	0.004
AP	/z/	ttip	372(255)	691(483)	1629(620)	951(473)	572(482)	E1/3	0.000	0.000
TP	/z/	ttip	2553(950)	4065(1798)	3801(1840)	1171(657)	1339(757)	E1/2	0.027	0.000
AP	/z/	tdor	233(149)	325(218)	444(219)	281(156)	216(88)	E1/3	0.011	0.002

Table A.13: Cases in which the pattern *decrease only* was found. Column 1: speaker, column 2: sound, column 3: sensor, columns 4-8: mean jerk values in different sessions, column 9: session with maximal jerk value, column 10: p of post-hoc test for the comparison session E1/1 - session with maximal value, column 11: p of post-hoc test for the comparison session with maximal value - session E3/1. For further details see section 11.2.

speaker	sound	sensor	E1/1	E1/2	E1/3	E2/1	E3/1	max	pI	pII
SK	/o/	tdor	1228(404)	1205(488)	1037(559)	431(258)	473(228)	E1/1		0.000
OP	/o/	jaw	3439(3773)		511(334)	265(96)	219(176)	E1/1		0.042
BP	/i/	ttip	525(479)	324(155)	390(209)	607(398)	296(132)	E2/1	1.000	0.038
DS	/i/	ttip	889(478)	1522(861)	1197(645)	566(266)	466(189)	E1/2	0.071	0.000
BP	/i/	tdor	440(292)	217(116)	412(197)	208(148)	195(127)	E1/1		0.020
BP	/t/	tdor	1677(720)	939(711)	1558(967)	1713(1133)	838(310)	E2/1	1.000	0.049
TP	/t/	tdor	7036(2708)	6939(2282)	9796(4392)	4855(2983)	5826(3978)	E1/3	0.200	0.047
BP	/z/	ttip	967(417)	991(649)	1214(641)	975(577)	703(350)	E1/3	0.828	0.032
TP	/z/	tdor	3079(1688)	1276(610)	1652(1047)	824(424)	885(478)	E1/1		0.000

Zusammenfassung

Die vorliegende Studie befasst sich mit der artikulatorischen Adaption als Folge einer Veränderung der Vokaltraktgeometrie durch einen künstlichen Gaumen. Ziel der Arbeit ist zu untersuchen, ob die Adaption darauf gerichtet ist, ein bestimmtes akustisches Resultat zu erreichen oder ob sie ein artikulatorisches Ziel hat. Dazu wurde ein Perturbationsexperiment durchgeführt, bei welchem die Vokaltraktgeometrie der Versuchspersonen durch eine Gaumenprothese verändert wurde, die von den Sprechern zwei Wochen lang getragen wurde.

Im weiteren Sinne untersucht die Studie phonemische Repräsentationen beim Sprecher. Konkreter und auf das Experiment bezogen ist die Fragestellung: Was ist das Ziel eines Adaptionsprozesses und was ist nur das Hilfsmittel? Streben Sprecher ein bestimmtes akustisches Resultat an und sind die artikulatorischen Bewegungen daher nur Hilfsmittel, oder versuchen die Sprecher, bestimmte artikulatorische Bewegungen zu produzieren und das akustische Resultat ist nur die Konsequenz dessen?

Eine Antwort auf diese Frage könnte neue Einblicke in die Mechanismen der Sprachproduktion geben. Je nachdem, mit welchem Ziel zu welchem Zeitpunkt adaptiert wird, kann man mit einem solchen Experiment feststellen, welche Charakteristika der Laute besonders wichtig sind, ob es eher akustische sind (z.B. bestimmte Frequenzbereiche) oder artikulatorische (z.B. ein bestimmtes linguo-palatales Kontaktmuster, ein bestimmter Konstriktionsort, eine Konstriktionsform). Die Resultate könnten das Verständnis für die Sprachproduktion allgemein erweitern, mehr Licht auf die Vorgänge bei der Sprache Gehörloser werfen oder Schwierigkeiten der Adaption beim Sprechen unter Perturbationen (z.B. Zahnprothesen) erklären.

Weiterhin betrifft die Frage nach der Art der Repräsentationen auch die Sprachperzeption, beispielsweise die Frage, ob Hörer direkt artikulatorische Aktivitäten oder Zustände wahrnehmen, oder ob sie die linguistische Information dem akustischen Signal entnehmen ohne dabei Rückschlüsse auf die artikulatorischen Vorgänge zu ziehen, die dieses Resultat produziert haben. Die perzeptuelle Frage, zu welcher die vorliegende Studie einen Beitrag leis-

ten könnte, ist - in anderen Worten - ob Sprecher, die einen Laut hören, direkt eine artikulatorische Aktion wahrnehmen, die durch ein akustisches Signal übertragen wird, oder ob sie die linguistische Information direkt dem akustischen Signal entnehmen und die artikulatorische Bewegung also nur ein Hilfsmittel ist.

Kapitel 1 der Studie diskutiert die bisherigen Untersuchungen und Ansätze zu artikulatorischen und akustischen Komponenten in den Phonemrepräsentationen sowohl in der Sprachproduktion als auch in der -perzeption. Zunächst erfolgt eine Begriffsdefinition der *speech production tasks* und der *perceptual primitives*. Der Begriff *perceptual primitives* bezeichnet in dieser Studie in etwa das, was Goldstein & Fowler (2003) *common currency* nennen. Es handelt sich dabei um die Information, die bei der Kommunikation übermittelt wird, also Sprecher und Hörer gemein ist. Im Unterschied dazu enthalten *speech production tasks* darüber hinaus noch motorische Informationen, die es dem Sprecher ermöglichen, effiziente Sprechbewegungen durchzuführen, die auch ohne auditive Rückmeldung möglich sind.

Nach dieser Begriffsklärung werden die beiden großen Richtungen in der Debatte um die Phonemrepräsentationen gegenübergestellt. Für die Richtung, die artikulatorische Ansätze vertritt, werden die *Motor Theory* und *Direct Realism* diskutiert. Einige Experimente, die diese Theorien unterstützen, werden beschrieben, zum Beispiel Cooper et al. (1952), die nachgewiesen haben, dass spektral gleiche Laute abhängig vom Kontext unterschiedlich wahrgenommen werden, was darauf zurückzuführen sein könnte, dass unterschiedliche artikulatorische Gesten wahrgenommen werden.

Anschließend wird ein beim Menschen möglicherweise vorhandener Mechanismus diskutiert, der gesturelle Wahrnehmung ermöglichen könnte: die Spiegelneuronen. Diese Neuronen, die beim Schimpansen gefunden wurden, sind aktiviert wenn das Tier eine Aktion ausführt, aber auch, wenn es beobachtet, wie jemand anders dieselbe Aktion ausführt. Diese Aktivierung kann deshalb als direkte Wahrnehmung eines motorischen Plans interpretiert werden, und die Spiegelneuronen könnten so die direkte Wahrnehmung artikulatorischer Information aus dem akustischen Signal ermöglichen.

Als akustische Theorien werden die *Theory of Acoustic Invariance*, die *Auditory Enhancement Theory* und die *Adaptive Variability Theory* diskutiert. Während die erste dieser Theorien davon ausgeht, dass jedes Phonem an einen fixen Bereich im akustischen Raum gekoppelt ist, nehmen die beiden anderen Theorien an, dass es Variabilität in der Akustik gibt. Die *Auditory Enhancement Theory* geht davon aus, dass es verschiedene akustische Parameter gibt, die wahlweise eingesetzt werden können und in der Perzeption ein und desselben Lautes resultieren können. Eine Untersuchung, die die *Auditory Enhancement Theory* stützt, ist die Studie von Diehl & Kingston(1991),

die sich mit verschiedenen akustischen Parametern, die alle in der Wahrnehmung des [voice]-Features resultieren, befasst. Die *Adaptive Variability Theory* geht davon aus, dass die separaten Regionen im akustischen Raum nur innerhalb einer Kommunikationssituation existieren und über die Situationen hinweg variieren können.

Kapitel 2 beschäftigt sich mit einem Aspekt der motorischen Komponente der *speech production tasks*, der Bewegungsoptimierung, und diskutiert frühere Untersuchungen dazu. Wie bei anderen Bewegungen auch, verfolgen Sprecher bei der Artikulation das Ziel, Bewegungen mit wenig Aufwand zu produzieren. Der Aufwand kann anhand verschiedener Parameter gemessen werden, zum Beispiel anhand des Rucks der Bewegung (engl. *jerk*, der dritte Ableitung der Bewegung, also der Beschleunigungsveränderung). Für das hier diskutierte Experiment kann man erwarten, dass die Sprecher, solange die Artikulation noch nicht perturbiert ist, Bewegungen mit geringem Ruck zeigen. Wenn die Artikulation perturbiert wird, werden sie vermutlich Bewegungen produzieren, die nicht optimal sind, weil sie zu viel Aufwand involvieren (größerer Ruck). Erst nach einer gewissen Adaptionszeit wird der artikulatorische Aufwand wieder sinken.

Den Abschluss des einführenden Teils der Arbeit bildet eine Diskussion früherer Perturbationsexperimente (Kapitel 3). Die Adaptionsstrategien, die in den Experimenten beobachtet werden, werden eingeteilt in Stabilisierung, Reparametrisierung und Reorganisation. Eine *Stabilisierung* erfolgt, wenn der Sprecher an einer erlernten Strategie festhält und versucht, diese Strategie zu stabilisieren, so dass die Perturbation keinen Einfluss auf die Artikulation hat. Diese Strategie wird z.B. angewandt, wenn Sprecher in verschiedenen Körperpositionen sprechen und daher unterschiedliche Gravitationskräfte auf Zunge und Kiefer wirken. *Reparametrisierung* ist eine Strategie, bei welcher die zu der Strategie gehörenden Komponenten die gleichen bleiben, sie aber eine andere Gewichtung erhalten. Bei Beißblockexperimenten etwa, übernehmen Zunge und Lippen Bewegungen, die normalerweise vom Kiefer ausgeführt werden. Eine *Reorganisierung* erfolgt, wenn die gesamte Strategie verändert wird, zum Beispiel wenn ein anderer Artikulator involviert ist oder sich die Vokaltraktform ändert. Das kann der Fall sein, wenn Sprecher eine z.B. tiefere Zungenposition beim /u/ durch mehr Lippenvorstülpung kompensieren. Für das hier durchgeführte Experiment werden über den Adaptionszeitraum hinweg verschiedene Strategien erwartet, zuerst eine Reparametrisierung, später eine Reorganisation.

Kapitel 4 beschreibt das Experiment und einige der Analysemethoden. Für jeden der sieben Sprecher wurde ein künstlicher Gaumen aus Acryl angefertigt, den die Versuchspersonen über einen Zeitraum von zwei Wochen ganztägig trugen. Die Versuchspersonen wurden aufgefordert, ihre Artiku-

lation so zu adaptieren, dass sie ihnen normal erscheint. Die artikulatorischen Bewegungen der Sprecher (Zunge, Kiefer und Lippen) wurden durch elektromagnetisch-artikulographische Aufnahmen (EMA) erfasst, die korrespondierenden akustischen Resultate wurden ebenfalls aufgezeichnet. EMA-Aufnahmen erfolgten zu Beginn des Experiments, nach einer Woche und am Ende des Experiments (nach zwei Wochen). Bei der ersten perturbierten Aufnahme wurde die auditive Rückmeldung durch die Darbietung von weißem Rauschen über Kopfhörer maskiert. Zusätzlich wurden nach 1.5 bzw 2.5 Wochen reine Audioaufnahmen unter Perturbation durchgeführt. Das Korpus bestand aus in Trägersätze eingebetteten Logatomen der Struktur $C_1V_1C_2V_2$. Sämtliche gespannten Vokale, ein ungespannter Vokal und alle lingualen Obstruenten des Deutschen wurden aufgenommen. Das akustische Signal wurde zeitlich in Einzellaute segmentiert. Für das artikulatorische Signal erfolgte eine Segmentation in artikulatorische Gesten. Außerdem wurden die artikulatorischen Zielpositionen der Laute gemessen. Die akustische Auswertung bestand aus Formantfrequenzmessungen für die Vokale und der Berechnung von sechs spektralen Parametern für die Frikative. Da es sich bei den akustischen Parametern um mehrere handelt, wurden Diskriminanzfunktionen berechnet mit Hilfe derer die Klassifizierbarkeit der Produktionen in Phoneme und die Qualitätsveränderungen der Produktionen über den Adaptionszeitraum untersucht wurden.

Die nachfolgenden Kapitel (Kapitel 5 bis 11) sind der Auswertung der Daten gewidmet. Kapitel 5 und das später folgende Kapitel 7 beschäftigen sich mit einer grundlegenden Bedingung für akustische Phonemrepräsentationen, der klaren Separierbarkeit der Phoneme im akustischen Raum zumindest innerhalb einer Aufnahme.

Kapitel 5 diskutiert die akustischen Charakteristika der Vokale, i.e. die Formantfrequenzen und die Dauern der Vokale. Zuerst wird untersucht, in wie weit die Frequenzen einzelner Formanten für die einzelnen Vokale divergieren. Nicht überraschend wird festgestellt, dass Vokale mit gleicher Zungenhöhe ähnliche F1 und Vokale mit gleicher Zungenlage ähnliche F2 haben. Zweidimensionale Regionen (F2-F1 und F2-F3) zeigen für alle Vokale Überlappungen, wenn die Produktionen aller Sitzungen insgesamt betrachtet werden. Wenn die Daten nach Sitzung aufgeteilt werden, sind die Überlappungen für die gespannten Vokale vernachlässigbar, die zwischen gespannten und ungespannten Vokalen bleiben beachtlich. Wenn die Dauer als Parameter hinzugezogen wird, wird die Unterscheidung nur in Ausnahmen besser. Eine Diskriminanzanalyse mit anschließender Klassifizierung in Phoneme in welche alle vier Parameter einflossen, bestätigt die Ergebnisse: Fast alle Produktionen können korrekt klassifiziert werden.

Die Klassifikationsergebnisse zeigen, dass die Separierung der Phoneme

in der ersten perturbierten Aufnahme schlechter ist als in der präperturbierten. In der Sitzung danach mit auditiver Rückmeldung wird die Klassifikation sogar noch schlechter. Die sehr niedrigen Klassifikationsraten in der Sitzung wenn auditive Rückmeldung verfügbar ist, werden damit begründet, dass die Variabilität der Produktionen bei einigen Sprechern in dieser Sitzung ansteigt. Das kann darauf zurückzuführen sein, dass die Sprecher ausprobieren, ob sie mit einer leicht veränderten Artikulation vielleicht zu einem besseren akustischen Resultat kommen können. Die Ergebnisse legen die Vermutung nahe, dass eine Klassifizierung anhand akustischer Parameter auf jeden Fall möglich ist. Informelle Perzeptionstests mit den Produktionen, die falsch klassifiziert wurden, lassen vermuten, dass diese Produktionen von Hörern richtig wahrgenommen werden. Das könnte bedeuten, dass nicht alle akustisch relevanten Parameter erfasst wurden. Ein zusätzlicher Parameter, insbesondere für die Distinktion gespannt-ungespannt könnte die Formanttrajektorie sein (Watson & Harrington (1999)).

Interessanter als die generell klaren Ergebnisse bei den Formantmessungen innerhalb einer Sitzung ist die Entwicklung über die Sitzungen hinweg. Über den gesamten Adaptionszeitraum verändern sich die Formantfrequenzen der Vokale leicht. Diese Verschiebungen erscheinen aber nicht zielgerichtet. Die Ergebnisse werden im Sinne der *Adaptive Variability Theory* interpretiert: In verschiedenen Kommunikationssituationen (verschiedenen Sitzungen) variieren die Formantfrequenzen leicht, innerhalb einer Sitzung sind sie jedoch fast eindeutig unterscheidbar.

Kapitel 6 beschäftigt sich nochmals mit der Vokalakustik und versucht, festzustellen, ob über die Adaption hinweg die artikulatorischen und akustischen Distanzen zwischen den Vokalen größer werden. Den Impuls zu dieser Untersuchung gab eine Entwicklung des F2 in /i/ und /e/ über den Adaptionszeitraum hinweg: F2 ist zu Beginn der Perturbation sehr viel niedriger als in der präperturbierten Aufnahme. Danach steigt er an. Die Resultate der artikulatorischen Analyse zeigen, dass die Zungenposition für die Hinterzungenvokale relativ stabil ist, während die Mehrheit der Sprecher die Zunge zu Beginn der Perturbation für die Produktion der Vorderzungenvokale zurückzieht und das Vokalviereck somit artikulatorisch verkleinert. Verbunden damit ist ein geringerer Abstand in einem akustischen Parameter (F2) zwischen Vorder- und Hinterzungenvokalen. Über den Adaptionszeitraum hinweg wird der Abstand zwischen Vorder- und Hinterzungenvokalen in beiden Bereichen, der Artikulation und der Akustik, größer. Wenn man die Resultate der Artikulation und Akustik vergleicht, kann man nicht sehen, dass die Veränderungen in einer Domäne stärker wären, also dass z.B. bestimmte akustisch stabile Regionen ausgenutzt werden und eigentlich nur die Artikulation variiert. Ein solches Resultat hätte natürlich stark für akustische

Komponenten in den Phonemrepräsentationen gesprochen.

Kapitel 7 beschäftigt sich mit der akustischen Adaption bei der Frikativproduktion. Dazu wurden sechs spektrale Parameter berechnet. Mit diesen Parametern wurden zwei Diskriminanzanalysen durchgeführt. Die erste Analyse hatte das Ziel, festzustellen, ob die Frikative innerhalb einer Teilaufnahme anhand akustischer Charakteristika voneinander unterscheidbar sind. Die Analyse zeigt, dass die Laute innerhalb einer Sitzung anhand akustischer Parameter gut klassifiziert werden können. Ähnlich wie die Vokale nehmen sie relativ gut separierbare Regionen im akustischen Raum ein. Wie bei den Vokalen verändern sich die Regionen über den Adaptionszeitraum hinweg. Eine zweite Diskriminanzfunktion wurde anhand der präperturbierten und der ersten perturbierten Aufnahme berechnet. Anschließend wurde die Wahrscheinlichkeit der Zugehörigkeit sämtlicher Produktionen zur Gruppe der präperturbierten Produktionen berechnet. Über den Adaptionszeitraum hinweg steigt diese Wahrscheinlichkeit. Daraus kann geschlussfolgert werden, dass, im Unterschied zu den Vokalen, die Veränderungen bei den Frikativen zielgerichtet sind: Die Produktionen werden akustisch gesehen den präperturbierten Produktionen immer ähnlicher. Die erste Analyse in diesem Kapitel zeigt also, dass eine Klassifizierung der Laute anhand akustischer Parameter möglich ist. Die zweite Analyse zeigt, dass es eine Entwicklung über den Adaptionszeitraum hinweg gibt, die sich auf akustische Parameter stützt. Die Analysen beschäftigen sich nicht mit den artikulatorischen Gegenstücken zu den akustischen Resultaten, diese Analyse wird auf Kapitel 9 verschoben.

Kapitel 8 beschäftigt sich, wie auch das darauf folgende Kapitel 9, mit dem Einfluss der auditiven Rückmeldung auf die Adaption. In Kapitel 8 wird die Vokalproduktion analysiert, Kapitel 9 ist den Frikativen gewidmet. Für die Vokalanalyse werden die Veränderungen im F1 in den ersten beiden perturbierten Aufnahmen (mit Maskierung der auditiven Rückmeldung und ohne) bei den Vokalen /i, e, o, u/ untersucht. Bei allen Vokalen ist der Unterschied im F1 zwischen der präperturbierten und der ersten perturbierten Sitzung (mit Maskierung der auditiven Rückmeldung) größer als zwischen der präperturbierten und der zweiten perturbierten Sitzung (ohne Maskierung). Die Verbesserung der Produktionen durch die auditive Rückmeldung ist jedoch größer für die obermittelhohen Vokale als für die hohen Vokale. Dieses Ergebnis kann mit der besseren Nutzung von taktiler Rückmeldung bei hohen Vokalen schon in der Sitzung ohne auditive Rückmeldung begründet werden. Die Analyse zeigt außerdem, dass Sprecher auch ohne auditive Rückmeldung mit Hilfe von taktiler Rückmeldung erstaunlich gut adaptieren können. Das bedeutet, dass die Sprecher über eine Art artikulatorische Repräsentation verfügen müssen, selbst wenn es vielleicht nicht die ist, die zur *common currency* gehört und an den Hörer übermittelt wird. Die artikula-

torische Repräsentation könnte ein Zunge-Gaumen-Kontaktmuster sein. Für andere denkbare artikulatorische Repräsentationen (artikulatorische Gesten) müsste man davon ausgehen, dass die Sprecher eine "Berechnung" ihrer Artikulationsbewegungen für die neue Umgebung durchführen können, nachdem sie, vor Beginn des Sprechens, die neue Umgebung durch taktile Rückmeldung erfasst haben.

Kapitel 9 vergleicht wie das vorangegangene Kapitel die ersten drei Teilaufnahmen (präperturbiert, perturbiert mit Maskierung der auditiven Rückmeldung, perturbiert ohne Maskierung), diesmal aber für die Frikative /s, ʃ, ç/ und /x/. Ziel der Auswertung ist wieder, den Einfluss der auditiven Rückmeldung auf das Adaptionsverhalten zu untersuchen. Dazu werden akustische und artikulatorische Analysen durchgeführt. Zuerst wird die Entwicklung des spektralen COG untersucht. Dabei wird festgestellt, dass der COG in der ersten perturbierten Sitzung fällt. Diese Veränderung wird in der Sitzung mit auditiver Rückmeldung aber nicht korrigiert, sondern der COG fällt weiter. Die Analyse artikulatorischer Zielpositionen unterstützt dieses Resultat. Die Sprecher verlagern die Zungenposition mit jeder Sitzung weiter nach hinten.

Eine Erklärung für diese Entwicklung, die das akustische Merkmal COG zu ignorieren scheint, kann mit Hilfe einer Analyse der Kieferposition in /s/ gefunden werden. Der Kiefer nimmt in der Sitzung mit Maskierung der auditiven Rückmeldung eine tiefere Position ein. In der darauf folgenden Sitzung wird der Kiefer wieder ein wenig angehoben. Dieses Ergebnis kann darauf zurückzuführen sein, dass die Sprecher in der Sitzung ohne Maskierung bemerken, dass die Amplitude des Rauschens im /s/ nicht mehr hoch genug ist, da die oberen Schneidezähne teilweise von der Prothese bedeckt werden, die unteren Schneidezähne durch die veränderte Kieferposition zu tief sind und daher der Luftstrom nicht auf das für die Sibilantproduktion notwendige zweite Hindernis trifft. Eine höhere Kieferposition ist wegen der Prothese jedoch bei gleicher Zungenlage nicht ohne Weiteres möglich, weil die Konstriktion sonst zu eng werden würde. Dies könnte der Grund für die Rückverlagerung der Zunge sein.

Es ist damit noch nicht klar, warum sich auch die artikulatorischen Positionen der Nicht-Sibilanten verändern. Die Veränderungen sind umso erstaunlicher als die Artikulation dieser Laute gar nicht direkt durch die Prothese, die nur den harten Gaumen bedeckt, beeinträchtigt wird. Eine Erklärung für die Veränderungen wäre eine Kettenverschiebung: Die Position des alveolaren Lautes wird verändert, um ihm die typischen Sibilantcharakteristika zu geben. Das mag auch beim postalveolaren Laut noch so sein. Der postalveolare Laut wird dadurch dem palatalen Frikativ so ähnlich, dass der palatale Frikativ "ausweicht", also weiter hinten produziert wird. Damit wird er dem velaren Frikativ sehr ähnlich, was zu einer Verschiebung der

velaren Artikulationsstelle führt. Die Analyse ist ein Beispiel dafür, dass artikulatorische Veränderungen stattfinden können, um akustische Ziele (eine bestimmte Amplitude der Frequenzen eines bestimmten Bereichs) zu erreichen. Unter Annahme der Adaption nach artikulatorischen Gesichtspunkten könnte ein solches Ergebnis nur begründet werden, indem man annimmt, dass die artikulatorische Repräsentation nicht nur aus der Information "alveolare Konstriktion" sondern auch noch aus der Information "zweites Hindernis im Luftstrom" besteht, welche die Sprecher in der Sitzung ohne Maskierung der Rückmeldung zu realisieren versuchen.

Kapitel 10 beschreibt einen sehr klassischen Test für akustische Phonemrepräsentationen. Es untersucht motorisch äquivalente artikulatorische Strategien in der Produktion des /u/, die alle zum gleichen akustischen Resultat führen. Um diesen Laut zu produzieren, können die Sprecher zum Beispiel die velare Konstriktion weiter oder enger wählen und dafür die Lippen mehr oder weniger vorstülpen. Für die Untersuchung an den hier diskutierten Daten wurden die velare Konstriktionsweite und die Lippenvorstülpung beim /u/ gemessen. Anschließend wurden Korrelationen zwischen beiden Parametern berechnet. Für die Mehrheit der Sprecher kann eine Kovariation zwischen Konstriktionsweite und Lippenvorstülpung tatsächlich gefunden werden. Dieses Resultat spricht für akustische Repräsentationen, da die Akustik des Lautes bei variierender Artikulation konstant gehalten wird.

Kapitel 11 beschäftigt sich ausschließlich mit der zweiten Komponente der *speech production tasks*, mit den motorisch motivierten Repräsentationen, und dort mit dem Teilaspekt der Optimierung der Bewegung. Dazu wurde der *Ruck*, der Aussagen über den artikulatorischen Aufwand ermöglicht, für die artikulatorischen Gesten einzelner Laute über die Adaptionszeit gemessen. Für die Zungenspitzenbewegung wurde festgestellt, dass bei Lauten, die leicht zu adaptieren sind (/o/), der Ruck über die Adaptionszeit zuerst ansteigt und dann abfällt. Das Ziel der Sprecher scheint daher, eine möglichst optimale Trajektorie zu finden. Um dieses Ziel zu erreichen, verändern die Sprecher die Artikulation leicht und probieren mehrere Strategien aus. Das resultiert in kleinen akustischen Unterschieden über die Aufnahmen hinweg. Kapitel 11 zeigt, dass die Optimierung der Bewegung aber zeitlich unabhängig vom akustischen Resultat ist. Es handelt sich also um einen Prozess, der parallel zu dem des Erreichens des akustischen Ziels erfolgt. Die Ergebnisse werden im Rahmen eines internen Models der Sprachproduktion im Sinne von Jordan (1996) interpretiert.

Kapitel 12 fasst die Studie zusammen und zieht Schlussfolgerungen. Einiges in den Daten spricht für akustische Komponenten in den *speech production tasks*. So gibt es, abgesehen von den ungespannten Vokalen, innerhalb einer Aufnahme relativ klar abgrenzbare Regionen im akustischen Raum für

jedes Phonem. Die Variation der akustischen Parameter über den Adaptionszeitraum kann im Sinne der *Adaptive Variability Theory* erklärt werden. Weitere Unterstützung für akustische Repräsentationen könnte man in der frühen Adaption des /s/ sehen, wo mehrere artikulatorischer Parameter (die Konstriktionsposition, aber auch die involvierten Artikulatoren und die Koordination von Zunge und Kiefer) variieren, um ein akustisches Ziel, nämlich ein hochfrequentes Geräusch zu erreichen. Ebenfalls für akustische Repräsentationen sprechen die motorisch äquivalenten artikulatorischen Strategien, die für das /u/ gefunden wurden.

Einige Ergebnisse sind aber mit rein akustischen Repräsentationen nicht zu vereinbaren. Die Sprecher adaptieren sehr erfolgreich, selbst wenn keine auditive Rückmeldung verfügbar ist. Die Tatsache, dass sie adaptieren, kann darauf zurückgeführt werden, dass sie die verfügbare taktile Rückmeldung nutzen, um eine vorhandene artikulatorische Repräsentation zu reparametrisieren. Es wird daher angenommen, dass eine artikulatorische Repräsentation mit motorischer Funktion existiert, also eine Repräsentation, die es nur beim Sprecher gibt, und die nicht zum Hörer übertragen wird. Die Existenz einer solchen Repräsentation könnte auch begründen, warum Sprecher, wenn ihre Artikulation nicht perturbiert ist, so selten Gebrauch von motorisch äquivalenten Strategien machen. Sie haben sich für eine Strategie entschieden (vermutlich weil es diejenige ist, die den geringsten Aufwand involviert) und verändern diese nicht mehr. Auch die Tatsache, dass Trade-offs, die in der Perzeption gefunden werden, oft nicht in der Produktion nachweisbar sind, jedenfalls nicht innerhalb der Sprache eines Sprechers, kann als Resultat der Auswahl einer optimalen Strategie gesehen werden, welche als artikulatorische Repräsentation fungiert. Gegen eine Funktion der artikulatorischen Repräsentationen als *perceptual primitives* spricht, dass die motorisch äquivalenten Strategien trotzdem existieren. Die eventuelle Existenz von Spiegelneuronen muss dieser Interpretation nicht entgegenstehen. Es ist möglich, dass wir motorische Pläne wahrnehmen können, dass sie aber nicht der linguistischen Information entsprechen.

Die vorgefundenen Adaptionsstrategien können als Reparametrisierung mit darauf folgender Reorganisation beschrieben werden. Am Beispiel der frühen /s/-Adaption kann man das gut demonstrieren. Zuerst wird der Kiefer zusammen mit der Zunge gesenkt und das Höhenverhältnis zwischen Zunge und Kiefer somit beibehalten. Erst später wird dieses Verhältnis verändert, es entsteht eine neue artikulatorische Strategie. Wie angenommen, kann die vollständige Adaption, wenn eine Reorganisation notwendig ist, sehr lange Zeit in Anspruch nehmen. Für das hier diskutierte Experiment kann man davon ausgehen, dass die Kompensation nach zwei Wochen nicht für alle Laute vollständig ist. Besonders für die Frikative ist mehr Zeit erforderlich.

Zusammenfassend lassen die Ergebnisse der Studie den Schluss zu, dass die Sprecher über eine akustische Repräsentation verfügen, die sie während der Adaption ansteuern. Darüber hinaus verfügen sie über eine artikulatorische Repräsentation, die motorische Funktion hat. Diese artikulatorische Repräsentation scheint reparametrisierbar zu sein und ermöglicht daher die ersten Adaptionsschritte wenn keine auditive Rückmeldung verfügbar ist. Später erfolgt eine Reorganisation, die zu einer Veränderung der artikulatorischen Repräsentation führt. Eine neue artikulatorische Repräsentation entsteht nach einem Optimierungsprozess, bei dem von verschiedenen motorisch äquivalenten Strategien eine ausgewählt wird.

Résumé

L'étude présentée ici s'intéresse aux stratégies d'adaptation de la production de la parole associées à un changement de la géométrie du conduit vocal dû à l'insertion d'une prothèse palatale. L'objectif ultime du travail est d'étudier si l'adaptation est guidée prioritairement par des buts acoustiques, par des buts articulatoires ou par une combinaison de buts de natures différentes. Pour cela une expérience de perturbation a été mise en place où la géométrie de la voûte palatine a été modifiée par une prothèse que les sujets de l'expérience ont porté pendant deux semaines.

Ainsi cette étude pose la question de la nature (articulatoire, acoustique ou multimodale) des représentations phonémiques chez le locuteur. Plus concrètement, et pour se concentrer sur l'expérience que nous avons élaborée, la question qui se pose est la suivante : quel est le but ultime du processus d'adaptation élaboré en réaction à la mise en place de la prothèse et quels sont les moyens mis en œuvre pour l'atteindre ? En d'autres termes, les locuteurs ont-ils pour but premier d'atteindre une cible acoustique, auquel cas les régularités que l'on pourrait observer dans le domaine articulatoire ne seraient qu'un moyen de réaliser cet objectif, ou les locuteurs essaient-ils de produire des mouvements articulatoires spécifiques, auquel cas les éventuelles régularités observées dans le domaine acoustique ne seraient qu'un effet secondaire de ces mouvements ? Répondre à cette question pourrait offrir des perspectives nouvelles sur les mécanismes sous-jacents à la production de la parole. En fonction des caractéristiques de la cible que les stratégies d'adaptation visent à atteindre, et de l'instant où cette cible peut être atteinte, on peut observer, dans le cas où la cible est acoustique, quelles sont les propriétés du signal de parole (par exemple certaines régions spectrales) qui sont spécialement liées à la réalisation de tel ou tel phonème, ou bien, dans le cas où la cible est articulatoire, quels sont les patrons articulatoires pertinents (contacts linguo-palataux, position de la constriction ou encore forme de la constriction). De tels résultats pourraient élargir la compréhension des mécanismes de la production de la parole en général, mais aussi celle des processus spécifiquement mis en jeu dans la production de la parole en conditions pa-

thologiques, comme dans le cas de locuteurs devenus sourds après la période d'acquisition de la parole, ou de ceux qui doivent gérer le port d'une prothèse dentaire.

La question de la nature des représentations des phonèmes ne se restreint pas au seul champ de la production de la parole. Elle renvoie bien-sûr tout autant à des questions fondamentales sur les mécanismes de perception de la parole, telles que celle de savoir si un locuteur perçoit directement des activités ou des états articulatoires, ou s'il retrouve l'information linguistique directement dans le signal acoustique sans qu'une référence aux mouvements articulatoires qui l'ont produit ne soit nécessaire. Plus concrètement, le débat auquel la présente étude souhaite contribuer dans le domaine de la perception de la parole, est celui du rôle du signal acoustique : l'auditeur n'utilise t'il ce signal que pour retrouver l'action articulatoire qui en est à l'origine, action articulatoire qui serait alors porteuse de l'information linguistique, ou retrouve t'il l'information linguistique directement dans les caractéristiques du signal acoustique, le mouvement articulatoire n'étant alors que le moyen de produire ces caractéristiques ?

Le chapitre 1 de ce manuscrit présente et discute les travaux de la littérature qui se sont intéressés aux composantes articulatoires et acoustiques des représentations phonémiques dans la production et la perception de la parole. Premièrement, les deux termes *tâche de production de la parole* et *primitives perceptives* sont définis. Le terme *primitives perceptives* sera utilisé pour qualifier ce que Goldstein and Fowler [2003] ont appelé *monnaie d'échange*. Il s'agit en l'occurrence de l'information qui est transmise du locuteur à l'auditeur au cours la communication parlée, et qui est donc partagée par les deux protagonistes de la communication parlée. Les *tâches de production de la parole* produisent cette information, mais elles ont de plus des propriétés caractéristiques qui ne sont pas liées à la tâche de la communication parlée, mais aux stratégies motrices sous-jacentes à la génération des mouvements des articulateurs qui pourraient en particulier permettre au locuteur d'optimiser ces mouvements.

Après cette définition des termes, les deux grands postulats théoriques, s'opposant sur le terrain de la caractérisation des représentations des phonèmes, sont présentés. En soutien au premier postulat, qui met en avant les aspects articulatoires, la *Théorie Motrice* et la théorie de la *Perception Directe* sont discutées. Quelques expériences, qui soutiennent ces théories, sont décrites. Citons parmi celles-là, les travaux de Cooper et al. [1952], qui ont prouvé que des sons de mêmes spectres pouvaient être perçus différemment selon contexte. Ce phénomène pourrait être dû, selon ces auteurs, à la perception des gestes articulatoires sous-jacents qui sont différents, plutôt qu'à celle des caractéristiques acoustiques similaires. De plus, un argument

proposé par Fowler [1996] est discuté. Cette dernière a en effet interprété la mutation phonétique diachronique qui a mené au développement du haut-allemand (Hochdeutsch), au cours de laquelle des plosives sont devenues des fricatives, comme un indicateur de la perception des gestes articulatoires aux dépens de la perception des sons. Elle s'est appuyée pour son argumentation sur le fait qu'au cours de cette mutation l'articulation a faiblement varié, alors que l'acoustique était radicalement transformée. Dans cette logique, il devrait y avoir eu une phase où les deux sons étaient perçus comme identiques, ce qui n'est possible que si la perception est celle des gestes articulatoires, seul élément quasi commun à ces deux réalisations.

Nous nous intéresserons ensuite à un mécanisme observé chez le singe, et susceptible d'exister chez l'homme, qui permet une association entre la perception de gestes et les commandes motrices permettant leur réalisation. Il s'agit des neurones miroirs. Ces neurones, qui ont été trouvés chez le chimpanzé, sont en effet actifs à la fois quand l'animal effectue une action dont l'intentionnalité est bien identifiée (prélèvement de nourriture, grimaces...), et quand il observe cette même action réalisée par un congénère ou un expérimentateur. Ce résultat peut être interprété comme la perception directe du plan moteur permettant la production du geste. Un tel mécanisme, s'il existe chez l'homme, et s'il s'applique à la perception auditive, pourrait donc permettre la perception directe du geste articulatoire dans le signal acoustique.

En soutien au deuxième postulat, qui met en avant les aspects acoustiques, la théorie de l'*Invariance acoustique*, la théorie du *Renforcement auditif* et la théorie de la *Variabilité adaptative* sont discutées. Alors que la première théorie suppose que chaque phonème est associé à une région spécifique et invariante de l'espace acoustique, les deux autres supposent qu'il existe de la variabilité dans l'acoustique, mais que cette variabilité n'empêche pas la récupération de l'information linguistique dans ce signal. La théorie du *Renforcement auditif* suppose qu'il existe plusieurs caractéristiques acoustiques qui peuvent être utilisées au choix et qui résultent toutes dans la perception du même phonème. Les travaux de Diehl and Kingston [1991] étayent cette théorie. Ils montrent l'existence de différentes variables acoustiques qui ne sont pas toujours toutes présentes mais qui peuvent toutes contribuer à la perception du trait de voisement. La théorie de la *Variabilité adaptative* suppose qu'il existe des régions séparées dans l'espace acoustique pour chaque situation de communication, et que ces régions se recouvrent d'une situation à l'autre. Des indices renseignant l'auditeur sur la situation de communication permettent de lever les ambiguïtés. Les travaux qui parlent en faveur de la perception de caractéristiques acoustiques, sont bien souvent des expériences qui perturbent l'articulation pour étudier les stratégies d'équivalence motrice qui produisent le même résultat acoustique, alors même que les locu-

teurs n'ont plus la possibilité de recourir aux gestes articulatoires habituels. Citons parmi ces travaux l'expérience des « tubes labiaux » (Savariaux et al. [1995], Savariaux et al. [1999]) où les lèvres de locuteurs sont maintenues artificiellement ouvertes lors de la production de la voyelle arrondie du français /u/. Il a été ainsi observé que ces locuteurs reculaient la langue, parce qu'ils étaient ainsi capables de produire sensiblement le même spectre acoustique qu'en condition normale, au prix d'une modification importante de l'articulation. Mais il existe aussi des exemples en parole non-perturbée tels que l'étude de Perkell et al. [1993], qui ont trouvé que, pour le /u/, les locuteurs jouaient en antagonisme sur la hauteur de la langue et sur le degré de protrusion labiale afin de préserver un résultat acoustique sensiblement constant.

Le chapitre 2 étudie un aspect de la composante motrice des *tâches de production de la parole*, l'optimisation des mouvements. Il discute des recherches passées sur ce thème. Comme cela a été montré pour d'autres mouvements humains, il est proposé que les locuteurs puissent s'efforcer de produire des mouvements tout en minimisant leur effort articulatoire. L'effort articulatoire peut être estimé à partir de plusieurs grandeurs physiques, telles que la durée, la vitesse, l'amplitude ou le jerk du mouvement. Au cours des mouvements, ces grandeurs ne peuvent cependant le plus souvent pas varier librement du fait de contraintes spécifiques. La durée du mouvement peut ainsi être imposée par la nature même de la tâche ; l'amplitude du mouvement peut être contrainte à rester à l'intérieur de limites données. La tâche peut donc imposer le type d'effort articulatoire susceptible d'être minimisé. Pour les gestes de pointage de cibles ou de séquences de cibles, il a été souvent proposé que l'optimisation gestuelle vise à obtenir les trajectoires les plus lisses possible. Un mouvement lisse est caractérisé par des variations lentes de l'accélération. Ceci correspond à une faible valeur du Jerk, la dérivée troisième du vecteur de position. A côté des travaux qui attribuent le caractère lisse des mouvement à la recherche d'un minimum d'effort (par exemple Jordan [1996]), d'autres propositions (Hogan [1984], Nelson [1983]) sont discutées qui supposent que de tels mouvements pourraient être privilégiés par les systèmes biologiques parce qu'ils empêcheraient que les mouvements n'atteignent les limites du système. Pour l'expérience discutée dans ce manuscrit, on peut donc s'attendre à ce que les locuteurs aient des mouvements lisses quand l'articulation n'est pas perturbée et que l'optimisation peut être mise en place. Ensuite, dans la première partie de la phase perturbée, sous l'effet des contraintes nouvelles introduites par la prothèse, il est vraisemblable que les mouvements ne seront pas optimisés parce que la compensation de la perturbation nécessite la découverte de stratégies nouvelles. Puis quand les locuteurs auront plus l'habitude de la prothèse, l'effort articulatoire pourrait

à nouveau baisser.

La fin de la partie introductive (chapitre 2 jusqu'au chapitre 3) est une discussion des expériences de perturbation publiées dans la littérature (chapitre 3). Les stratégies d'adaptation qui ont été observées au cours de ces expériences, sont classifiées en « stabilisation », « reparamétrisation » et « réorganisation ». Une *stabilisation* s'effectue quand le locuteur s'efforce de conserver une stratégie articulatoire acquise et de la rendre robuste, pour que la perturbation n'ait pas d'influence sur la configuration articulatoire finale. Cette stratégie est utilisée par exemple quand un locuteur parle dans des positions diverses du corps et qu'en conséquence il a à faire face à des forces externes variables qui s'exercent sur la langue et sur la mandibule. La *reparamétrisation* est une stratégie où les différentes composantes impliquées dans la production du mouvement restent les mêmes, mais où leurs contributions respectives dans l'accomplissement de la tâche varient en fonction des conditions. C'est ce qui se passe par exemple au cours des expériences de perturbation de la position mandibulaire avec un bite-block (petit cube rigide inséré entre les dents et maintenant la mandibule à une position constante), pour lesquelles on observe que la langue et les lèvres prennent en charge les déplacements qui sont normalement produits par la mandibule. Bien que les stratégies articulatoires changent sous l'influence de la perturbation, les caractéristiques fondamentales du mouvement de l'effecteur final restent les mêmes. La caractéristique géométrique cible (par exemple la position de la constriction dans le conduit vocal) reste ainsi sensiblement inchangée. On est en présence d'une *réorganisation* quand les propriétés plus fondamentales de la stratégie motrice sont modifiées. C'est le cas par exemple quand une constriction est formée à un endroit radicalement différent du conduit vocal, ou quand un locuteur produit une protrusion exagérée des lèvres pour compenser une perturbation intervenant dans le conduit vocal. Dans le cas spécifique de l'expérience présentée dans ce manuscrit on peut s'attendre, compte tenu de la nature de la perturbation choisie, à observer une reparamétrisation dans une première phase puis une réorganisation des stratégies dans une seconde phase.

Le chapitre 4 décrit l'expérience et la méthodologie d'analyse exploitées dans ce travail. Pour chacun des sept locuteurs une prothèse palatale spécifique a été réalisée en matériau acrylique. Elle a été portée par chacun d'entre eux durant la journée pendant deux semaines. Il a été demandé aux sujets d'adapter leur articulation pour que leur production de la parole paraisse normale. Différentes séances d'enregistrement ont eu lieu pour chaque locuteur au cours de ces deux semaines. Lors de la première séance, les mouvements des articulateurs (langue, mandibule, lèvres) ont ainsi été enregistrés par articulographie électromagnétique (EMA) en synchronie avec le signal

acoustique. Cette séance a consisté en trois sessions. Au cours de la première d'entre elles, les locuteurs ont été enregistrés sans la prothèse ; cet enregistrement sert de référence. Pendant la seconde session, les sujets portaient la prothèse et leur feedback auditif était masqué par du bruit blanc perçu via un casque. Enfin la troisième session s'est effectuée avec la prothèse et avec un retour auditif normal. Les mêmes données (articulatoires et acoustiques) ont été recueillies lors de la dernière séance d'enregistrement (après donc deux semaines) qui était composée de deux sessions. La première d'entre elles a correspondu à des enregistrements avec prothèse, et la seconde à des enregistrements en condition normale. Cette dernière session avait comme objectif de permettre d'étudier les répercussions immédiates des processus d'adaptation mis en œuvre au cours des deux semaines passées sur la parole de nouveau non-perturbée. Outre ces trois séances, deux séances intermédiaires ont permis les enregistrements audio (sans EMA) de nos sujets, en condition perturbée, 1.5 et 2.5 semaines après le début de la période d'étude. Le matériau linguistique était composé de logatomes de structure $C_1V_1C_2V_2$ qui étaient insérés dans une phrase porteuse. Les sons étudiés étaient en position C_1 ou V_1, à trois exceptions près /s/, /ç/ et /x/. Ces sons ne sont en effet jamais en position initiale en allemand et ils ont donc été enregistrés en position C_2. Toutes les voyelles tendues, une voyelle relâchée, et toutes les fricatives et plosives non labiales ont constitué notre corpus. Le signal acoustique a été segmenté en sons élémentaires. Le signal articulatoire a été segmenté en gestes élémentaires à partir des passages par zéro de la vitesse tangentielle. De plus, les positions cibles articulatoires associées à chaque son élémentaire ont été mesurées. Les fréquences de formants ont calculées pour les voyelles. Pour la caractérisation spectrale des fricatives, les valeurs du centre de gravité (COG), de l'étalement fréquentiel (dispersion), de la dissymétrie (skewness) et de la proéminence (kurtosis) ainsi que celles des pentes approximant l'enveloppe spectrale entre 700 Hz et 2.5 kHz (a-slope) et entre 2.5 kHz et 12 kHz (b-slope) ont été calculées. Dans cet espace multidimensionnel, une fonction discriminante a été calculée qui a permis de voir dans quelle mesure les productions ainsi caractérisées sont classifiables en phonèmes et si la distribution de ces classes évolue pendant la phase d'adaptation.

Les chapitres qui suivent (chapitres 5 jusqu'à 11) sont dédiés à l'analyse des données. Les chapitres 5 et 7 s'intéressent à la question de savoir si les productions peuvent être classifiés par phonème, sur la seule base de la caractérisation acoustique. Une telle classification est en effet une condition nécessaire à la défense des théories soutenant la prédominance des représentations acoustiques. Cependant, les théories qui prônent l'existence de cibles articulatoires ne nient pas le fait que des régions phonémiques invariantes puissent

exister dans le signal acoustique, en tant que conséquences des invariants articulatoires. L'existence de régions distinctes dans l'espace acoustique n'est donc pas à elle seule la preuve de l'existence de représentations acoustiques pour les phonèmes.

En ce qui concerne les voyelles (chapitre 5), les locuteurs ont en général qualifié de « non-problématique » l'adaptation de leur articulation à la prothèse. Dans un premier temps, la variabilité des fréquences formantiques des voyelles isolées est étudiée. Conformément aux connaissances phonétiques classiques, on observe que les voyelles ayant une même hauteur de la langue ont des F1 similaires, et que celles qui ont une même position horizontale de la langue ont des F2 similaires. Si on analyse ces données dans les plans (F2-F1) et (F2-F3) on constate par conséquent de nombreux chevauchements entre les zones de réalisation des différents phonèmes. Même si on prend en plus en compte la durée, les effets sont faibles, et une séparation entre les phonèmes n'apparaît que dans quelques cas exceptionnels. Ainsi donc dans nos données, aucune tendance à une compensation entre durée et fréquences formantiques ne semble pouvoir s'opérer dans le domaine perceptif. Cependant, si on analyse les données session par session, le degré de chevauchement entre phonèmes devient négligeable dans le cas des voyelles tendues. Dans ce cas en effet, une analyse discriminante dans l'espace des quatre paramètres acoustiques (formants et durée) mentionnés ci-dessus permet une bonne classification en phonèmes.

Au-delà de ces résultats qui montrent le maintien d'une bonne distinction entre phonèmes pour chaque session prise séparément, l'analyse de l'évolution des caractéristiques acoustiques à travers les sessions successives, au cours de la phase d'adaptation, est particulièrement intéressante. On peut en effet observer que les zones formantiques associées à chaque phonème varient légèrement d'une session à l'autre. A deux exceptions près, ces changements ne paraissent cependant pas dirigés vers un but précis. Ainsi, on n'observe pas chez nos locuteurs de tendance à produire des formants qui seraient de plus en plus proches des formants mesurés au cours de la session non perturbée initiale, qui sert de référence pour caractériser la parole normale. Ce résultat peut être interprété dans le contexte de la théorie de la *Variabilité adaptative* : d'une condition de communication à l'autre (en l'occurrence pour nous, d'une session à l'autre) les fréquences de formants varient, ce qui induit des recouvrements entre des phonèmes produits dans des sessions différentes, mais ceci n'empêche pas, nous l'avons dit ci-dessus, le maintien de la distinction à l'intérieur d'une session donnée. Plus encore, des tests de perception informels sur les productions qui étaient classifiées de manière erronée par l'analyse discriminante, ont même laissé supposer que ces productions pourraient être perçues correctement. Ce résultat suggère donc que les quatre

variables choisies n'étaient pas les seuls paramètres utiles à une bonne classification acoustique. Une autre caractéristique pourrait être la trajectoire des formants, en particulier pour la distinction tendu/relâché (Watson and Harrington [1999]).

Une autre évolution intéressante au cours de la phase d'adaptation est celle des résultats de la classification en phonèmes. Cette classification était moins bonne dans la première session perturbée, quand le feedback auditif était masqué, que dans la session non-perturbée qui a précédé. Ceci est conforme à nos attentes. Mais, au cours de la session suivante qui ne perturbait pas le feedback auditif, la classification s'est encore dégradée. Ce n'est qu'après une demi-semaine d'adaptation que ces valeurs sont sensiblement revenues au niveau de celles de la session non-perturbée initiale. La mauvaise classification obtenue sur les sons prononcés lorsque le feedback auditif était pour la première fois disponible après la modification de la forme du palais, peut être expliquée par la grande variabilité des réalisations de quelques locuteurs au cours de cette session. Cette grande variabilité pourrait avoir pour origine le fait que les locuteurs aient testé plusieurs stratégies articulatoires, vraisemblablement dans le but d'obtenir un meilleur résultat acoustique, mettant à profit le fait que le feedback auditif, de nouveau disponible, leur permettait d'en contrôler directement l'efficacité. Ces résultats soutiennent l'hypothèse qu'une classification des phonèmes sur des critères purement acoustiques est possible.

Le chapitre 6 s'intéresse encore une fois à la production des voyelles, cette fois-ci sous l'angle d'une analyse détaillée des changements opérés pendant la phase d'adaptation. Plus concrètement, il s'agissait de voir si, pendant la phase d'adaptation, les distances articulatoires et acoustiques entre les voyelles devenaient plus grandes. Cette analyse a été stimulée par l'observation de la variation du F2 du /i/ et du /e/ pendant la phase d'adaptation. Au cours de la première session perturbée, ce formant était en effet moins élevé pour ces deux voyelles que pendant la session non-perturbée initiale. Les résultats de l'analyse articulatoire montrent que les positions de la langue pour les voyelles d'arrière sont restées relativement stables, tandis que pour les voyelles d'avant la majorité des locuteurs reculaient la langue, produisant ainsi un quadrilatère articulatoire des voyelles plus petit, caractérisé par une distance plus faible entre voyelles d'arrière et voyelles d'avant. Ceci est conforme à la diminution de la distance observée selon la dimension F2. Dans les sessions suivantes de la phase d'adaptation, cette distance est redevenue constamment plus grande dans les deux domaines, dans le domaine articulatoire et dans l'acoustique. Il semble donc que les locuteurs aient mis à profit la période d'adaptation pour augmenter la distance entre le groupe des voyelles d'avant et celui des voyelles d'arrière. Il n'est cependant pas certain

que cette augmentation ait été pertinente du point de vue perceptif, puisqu'il y a rarement ambiguïté perceptive entre voyelles d'arrière et voyelles d'avant. Les contrastes perceptifs plus sensibles, comme par exemple celui entre /i/ et /e/ n'ont par contre pas augmenté au cours des différentes sessions, ni dans le domaine de l'articulation (hauteur de la langue) ni dans celui de l'acoustique (F1). Il est cependant possible que les locuteurs se soient efforcés de maximiser de manière générale les distances, mais que ceci ne se soit pas avéré possible partout dans l'espace acoustique. On pourrait concevoir cette stratégie comme un mécanisme général qui s'applique à l'ensemble des voyelles, même si le contraste entre les phonèmes n'est pas menacé. Pour les deux voyelles produites à l'avant il est possible qu'il soit tout simplement impossible, dans les conditions imposées par la perturbation d'augmenter leur contraste. L'augmentation générale de la taille de l'espace vocalique serait la trace des efforts mis en place, vainement, par les locuteurs. Une autre explication pourrait aussi être que la perturbation a influencé de manière significative la position horizontale de la langue seulement pour les voyelles de /i/ et /e/. Ceci pourrait être envisageable si on considère que les locuteurs ont simplement essayé d'éviter le contact avec la prothèse et ont de ce fait initialement changé la position de la langue seulement pour les voyelles d'avant directement perturbées mécaniquement par la prothèse. Un tel recul pour ces deux voyelles ne les mettrait pas en situation d'être confondues avec d'autres puisqu'il n'y a phonologiquement aucune voyelle centrale dans le système allemand. Par la suite l'articulation se serait réajustée sur la base des seules habitudes articulatoires tant que le contact avec la prothèse n'a pas empêché l'antériorisation du positionnement lingual. Quand on compare les résultats de l'articulation avec ceux de l'acoustique on ne note pas des changements plus forts dans un domaine que dans l'autre. Ainsi donc on ne peut pas conclure que les changements se soient produits à l'intérieur de régions acoustiques peu sensibles aux variations articulatoires, résultat qui aurait été en faveur de l'hypothèse de représentations dans le domaine de l'acoustique.

Le chapitre 7 est consacré à l'étude de l'adaptation dans le domaine acoustique pour les fricatives. Dans ce but les six variables spectrales mentionnées plus en haut ont été calculées. Sur cette base, deux analyses discriminantes ont été effectuées. La première avait comme but d'étudier si, à l'intérieur d'une même session, les fricatives pouvaient être distinguées à partir des seules caractéristiques acoustiques. L'analyse a montré que c'était effectivement le cas, de manière générale. Comme les voyelles, les fricatives ont occupé des régions différentes dans l'espace acoustique. Et, comme pour les voyelles, les régions ont aussi changé d'une session à l'autre. Une deuxième analyse discriminante, qui avait comme but de caractériser le développement

de la qualité des productions à travers les différentes sessions, a été conduite avec comme classes-références la session non-perturbée initiale et la première session perturbée (sans feedback auditif). Sur cette base, la probabilité des productions des sessions suivantes de faire partie de la classe « session non-perturbée » ou de la classe « session perturbée initiale » a été calculée. Au cours de la période d'adaptation la probabilité d'être classifiée « session non-perturbée » a augmenté. Les productions sont donc devenues plus proches des productions normales. On peut donc conclure que, contrairement à ce qui a été observé pour les voyelles, les changements opérés pour les fricatives pourraient avoir un but précis, celui de produire des sons qui sont acoustiquement similaires aux productions non-perturbées. La première analyse montre donc qu'une classification des sons à partir des paramètres acoustiques est possible. La deuxième analyse montre qu'il y a un développement pendant l'adaptation qui s'appuie sur des paramètres acoustiques pour se rapprocher des productions normales. Les analyses présentées dans ce chapitre ne traitent pas des pendants articulatoires des résultats acoustiques ; ces résultats seront présentés dans le chapitre 9.

Le chapitre 8 s'intéresse à l'influence du feedback acoustique sur l'adaptation pour les voyelles. Dans ce but les changements de F1 dans les deux premières sessions perturbées (avec masquage du feedback auditif et sans masquage) pour les voyelles /i, e, y, o, u/ ont été mesurés. Pour toutes ces voyelles, la différence selon F1 entre la session non-perturbée initiale et la première session perturbée (avec masquage auditif) est plus grande que celle qui existe entre la session non-perturbée initiale et la deuxième session perturbée (sans masquage auditif). Cette différence n'est toutefois pas toujours significative. L'amélioration dans la session avec feedback auditif semble être plus importante pour les voyelles les plus basses (/e, o/) que pour les voyelles les plus hautes /i, u, y/. Certes cette différence n'est pas statistiquement significative, mais la tendance existe et ce résultat pourrait être expliqué par une meilleure exploitation du feedback tactile pour les voyelles hautes, lorsque le feedback auditif n'est pas présent. L'analyse montre en effet que les locuteurs peuvent compenser la perturbation même sans feedback auditif. Cela suggère que les locuteurs pourraient disposer d'une représentation articulatoire de la tâche. Cette représentation articulatoire pourrait être un patron de contacts linguo-palataux. Si on imagine d'autres types de représentation dans le même domaine (par exemple sous forme de gestes articulatoires), on doit supposer que les locuteurs seraient capables de faire une prédiction, dans le nouvel environnement, des conséquences du mouvement articulatoire, après avoir pris connaissance de ce nouvel environnement par feedback tactile avant de commencer à parler. La question reste alors posée de savoir si cette représentation articulatoire, quelle que soit sa forme, fait partie intégrante de la

monnaie d'échange qui est transmise à l'auditeur.

Le chapitre 9 étudie, comme le chapitre précédent, l'évolution des production au cours des trois premières sessions, mais cette fois-ci pour les fricatives /s, ʃ, ç/ et /x/. Le but de cette analyse est encore une fois d'étudier l'influence du feedback auditif sur le comportement au cours de la période d'adaptation. Des analyses acoustiques et articulatoires ont été effectuées. Premièrement le développement du COG spectral a été mesuré. Les résultats montrent que le COG baisse dans la première session perturbée, et que ce changement, loin d'être corrigé dans la session suivante avec feedback auditif, se poursuit dans les deux sessions suivantes. L'analyse des positions articulatoires cibles va dans le sens de ce résultat : les locuteurs reculent en effet toujours plus leur langue d'une session à l'autre. Une explication pour ce développement, qui semble d'ignorer le paramètre acoustique COG, peut être trouvé à l'aide d'une analyse des positions de la mandibule pendant la production du /s/. Dans la session avec masquage du feedback auditif, la mandibule prend une position assez basse. Dans la session qui suit, avec feedback auditif, la mandibule a une position un peu plus haute, plus proche de la position originale observée dans la session non-perturbée initiale. Une explication pour ce résultat pourrait être que, dans la troisième session, quand le feedback auditif est de retour, les locuteurs remarquent que l'amplitude du bruit pendant la production de /s/ n'est pas suffisamment forte. Cette baisse d'intensité serait la conséquence du fait que les incisives supérieures sont partiellement couvertes par la prothèse et que les incisives inférieures sont trop basses, en raison de la position ouverte de la mandibule. En conséquence l'écoulement d'air pourrait ne pas trouver l'obstacle qui est nécessaire à la production du bruit pendant les sons sibilants en allemand. Pour corriger cela, les locuteurs pourraient choisir d'adopter une position plus haute de la mandibule, qui permettrait d'utiliser les incisives inférieures comme obstacle. Sans geste compensatoire de la langue, cette élévation mandibulaire induirait une forte diminution de la taille de la constriction. C'est pourquoi, selon cette perspective, les locuteurs adopteraient une position de la langue plus reculée, responsable de l'abaissement observé pour le COG.

Cependant, si cette explication justifie les observations faites pour le /s/, elle ne nous éclaire pas sur les raisons pour lesquelles les positions articulatoires des non-sibilantes sont aussi reculées, puisque ces sons n'ont pas besoin d'un obstacle pour leur bonne production acoustique. En réalité, pour ces sons, ces changements de position sont d'autant plus surprenants que ces sons ne sont pas directement perturbés par la prothèse qui ne couvre que le palais dur. Ces changements pourraient alors s'expliquer par un mécanisme de réactions en chaîne. La position du son alvéolaire serait donc changée, ainsi que nous le suggérons, afin de préserver les caractéristiques

acoustiques d'une sibilante. Cette hypothèse pourrait aussi s'appliquer au son post-alvéolaire. Ce recul rendrait le son post-alvéolaire trop similaire de la fricative palatale. Pour préserver une différence entre ces deux sons, un recul s'opérerait alors aussi pour la fricative palatale, ce qui la rendrait trop similaire de la fricative vélaire, pour laquelle une articulation plus postérieure serait alors adoptée afin de retrouver un contraste suffisant. Cette analyse offre un cadre cohérent pour comprendre comment le changement articulatoire pourrait être élaboré et propagé, dans le but initial d'atteindre une cible acoustique, en l'occurrence une énergie du bruit consonantique suffisante dans les fréquences hautes, pour la fricative alvéolaire, puis dans l'objectif de maintenir un contraste acoustique suffisant entre les fricatives dans une zone globalement déplacée. Trouver une explication purement articulatoire aux changements observés semble être plus difficile. Il faudrait en effet supposer que la représentation articulatoire ne consiste pas seulement en une information du type « constriction alvéolaire », mais aussi en une information du type « générer un obstacle dans l'écoulement de l'air » pour laquelle la question des afférences oro-sensorielles serait posée.

Le chapitre 10 décrit un test très classique pour la perception des caractéristiques acoustiques sans référence à l'articulation. Il étudie les différentes stratégies articulatoires observées pour un même résultat acoustique lors de la production du /u/. On y constate que les locuteurs peuvent utiliser des stratégies diverses pour atteindre un même but acoustique. La variabilité articulatoire est ainsi tolérée, pourvu que le résultat acoustique reste le même. Pour l'analyse de ces stratégies d'équivalence motrice, on a testé si la largeur de la constriction et le degré de protrusion labiale étaient corrélés. L'hypothèse sous-jacente à cette démarche est que, si les locuteurs ont des cibles acoustiques comme représentations des phonèmes, ils pourraient utiliser la relation de compensation entre la largeur de la constriction et le degré de protrusion des lèvres pour trouver une stratégie robuste de production de la voyelle en présence de la perturbation palatale. On peut, par exemple imaginer la production d'une plus forte protrusion labiale associée à une constriction plus grande afin de ne pas risquer de produire une occlusion du conduit vocal quand la prothèse est installée. Notre analyse montre qu'en effet, pour la majorité des locuteurs, une corrélation significative existe entre la protrusion des lèvres et la largeur de la constriction. Ce résultat va clairement dans le sens de l'hypothèse des représentations acoustiques puisque tout semble fait pour maintenir les propriétés acoustiques des sons intactes, alors même que l'articulation varie.

Le chapitre 11 traite exclusivement de la deuxième composante, celles des *tâches de production de la parole*, sous l'angle de la mise en exergue de propriétés articulatoires qui pourraient être motivées par des raisons mo-

trices, telles que l'optimisation du mouvement. Pour cela, le *jerk*, qui permet d'estimer l'effort articulatoire, a été calculé pour les gestes à l'origine de la production des différents sons, et ceci pour la session non-perturbée initiale et pour toutes les sessions perturbées. Pour la pointe de la langue et pour les sons pour lesquels il a été facile de compenser très vite l'impact de la prothèse palatale (par exemple pour le /o/), le jerk augmente dans les sessions suivant immédiatement la mise en place de la perturbation, avant de baisser dans les sessions situées à la fin de la période d'adaptation. Le but des locuteurs semble donc être de produire des mouvements optimaux, associée à la production d'un minimum d'effort. Pour y arriver, les locuteurs changent l'articulation légèrement et essaient plusieurs stratégies, ce qui a des incidences acoustiques mineures telles que celles qui ont été observées sur les formants des voyelles. L'analyse montre aussi que l'optimisation du mouvement a été temporellement indépendante du résultat acoustique. On ne constate en effet pas que la recherche d'une stratégie optimale ait attendu que la cible acoustique ait été atteinte pour débuter. Donc, c'est un processus purement moteur qui semble s'être déroulé parallèlement à celui qui a eu pour but d'atteindre des objectifs acoustiques porteurs de l'information linguistique. Les résultats sont interprétés dans le cadre de l'hypothèse de l'existence d'un modèle interne de la production de la parole conforme aux propositions de Jordan [1996].

Le chapitre 12 résume notre étude et propose plusieurs conclusions. Certains de nos résultats corroborent l'hypothèse de l'existence d'une composante acoustique dans la spécification des objectifs des *tâches de production de la parole*. Il y a, si on fait abstraction du contraste voyelles tendues versus voyelles relâchées, des régions clairement séparées pour les phonèmes dans l'espace acoustique. La variation des caractéristiques acoustiques pendant la phase d'adaptation peut-être interprétée dans le cadre de la théorie de la *Variabilité adaptative*. L'hypothèse des représentations acoustiques est aussi renforcée par l'observation des stratégies d'adaptation développées pour le /s/ lors de la première séance d'enregistrement (sessions 1 à 3), au cours de laquelle plusieurs paramètres articulatoires (position de la constriction, mais nature des articulateurs recrutés pour ce geste et coordination entre langue et mandibule) varient d'une façon cohérente avec le but d'atteindre une cible acoustique, en l'occurrence le bruit caractéristique d'une sibilante alvéolaire. Des stratégies d'équivalence motrice, observées en particulier pour la production du /u/ soutiennent aussi l'hypothèse de l'existence de représentations acoustiques.

Par ailleurs, il y a des résultats qui ne sont pas compatibles avec l'idée de représentations purement acoustiques. Les locuteurs élaborent en effet des stratégies d'adaptation efficaces, même en l'absence de feedback auditif. On

peut attribuer ce phénomène à la prise en compte du feedback tactile que les locuteurs pourraient utiliser pour une reparamétrisation de leurs stratégies articulatoires, feedback tactile qui ferait ainsi partie de la représentation articulatoire du son. Il est donc probable qu'il y ait une représentation articulatoire de la tâche de parole, mais nous pensons qu'il s'agit d'une représentation associée à une fonction motrice, une représentation qui n'existe que pour le locuteur, afin de l'aider à l'accomplissement de sa tâche motrice, et qui n'aurait pas pour but d'être transmise à l'auditeur à des fins linguistiques. L'existence d'une telle représentation pourrait aussi expliquer pourquoi en parole non-perturbée on observe assez rarement des stratégies d'équivalence motrice. Il semblerait que les locuteurs se décident pour une certaine stratégie (peut-être parce que c'est celle qui correspond à un minimum d'effort) et qu'ils n'en changent plus tant que des perturbations externes ne rendent pas son exécution impossible. Cette sélection d'une stratégie unique pourrait expliquer pourquoi pour un locuteur donné on n'observe pas toutes les stratégies qui seraient compatibles avec une bonne perception du son produit. Nos locuteurs ont une représentation articulatoire correspondant à la stratégie impliquant le moins d'effort. Un argument contre un possible rôle de ces représentations articulatoires dans la caractérisation des *perceptual primitives* est justement le fait que des stratégies d'équivalence motrice existent, qui préservent les caractéristiques perceptives du son produit, et ceci aux dépens de la constance de la stratégie articulatoire.

Les stratégies d'adaptation observées ici peuvent être classifiées comme une reparamétrisation suivie d'une réorganisation. On peut bien le montrer pour l'adaptation du /s/. Premièrement la mandibule et la langue sont abaissées à cause de la prothèse palatale, et la relation d'hauteur entre mandibule et langue est maintenue. Cette relation est modifiée plus tard, et les locuteurs trouvent une nouvelle stratégie articulatoire. Comme prévu, l'adaptation se poursuit avec une réorganisation qui peut prendre beaucoup de temps. Pour l'expérience discutée ici, l'adaptation n'était probablement pas finie pour tous les sons, en particulier pour les fricatives pour lesquelles plus de temps aurait été nécessaire.

En conclusion, les résultats de notre étude suggèrent que les locuteurs ont une représentation acoustique des phonèmes, et que cette représentation est le but qu'ils essaient d'atteindre pendant l'adaptation et qu'ils veulent transmettre à l'auditeur. Les locuteurs disposent en outre d'une représentation articulatoire qui a une fonction purement motrice. Cette représentation semble être reparamétrisable et elle permet donc une adaptation immédiate même si il n'y a pas de feedback auditif. Par la suite, une réorganisation s'opère, qui amène un changement de la représentation articulatoire. Une nouvelle représentation articulatoire se développe après un processus d'optimisation,

au cours duquel une stratégie est élaborée via un choix parmi de nombreuses autres stratégies motrices équivalentes sur le plan de leurs conséquences sur l'acoustique.