Proceedings of Meetings on Acoustics

Volume 9, 2010 http://asa.aip.org

159th Meeting
Acoustical Society of America/NOISE-CON 2010
Baltimore, Maryland
19 - 23 April 2010
Session 2pSC: Speech Communication

2pSC19. Dependency of compensatory strategies on the shape of the vocal tract during speech perturbed with an artificial palate

Jana Brunner*, Philip Hoole, Frank Guenther and Joseph S. Perkell

***Corresponding author's address: Speech Communication Group, Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, Massachusetts 02139, jbrunner@mit.edu**

This study explores the idea that a speaker's choice of a strategy to compensate for a vocal-tract perturbation depends on the shape of the perturbed vocal tract. Speakers' palatal shapes were perturbed with palatal prostheses. Three speakers used an alveolar prosthesis that effectively moved the alveolar ridge toward the back; three used a central prosthesis that effectively flattened the palate. We hypothesized that during production of the front-rounded vowel /y/ the speakers with the alveolar prosthesis would compensate for the shortened anterior cavity with increased lip protrusion. Lip and tongue movement data from EMA recordings of the speakers' adaptive behavior supported the hypothesis: those whose front cavity was shortened by the palatal prosthesis increased lip protrusion; those with a flattened palate did not. This difference in adaptation strategies was investigated further using simulations with the DIVA model of speech production. The model's vocal tract was adapted to fit two of the speakers' vocal tracts (one with each type of prosthesis), using vocal-tract shape data from structural MRI recordings. Simulations of the model agree with the experimental results: compensation for the alveolar prosthesis was accomplished mainly with lip protrusion, whereas with the central prosthesis, it was accomplished with tongue movement.

Published by the Acoustical Society of America through the American Institute of Physics

Received 4 Jun 2010; published 16 Jun 2010

1. Introduction

The German front-rounded vowel /y/ is characterized by a palatal constriction and lip protrusion. The acoustics of this sound are determined mainly by the respective lengths of the front cavity (F2) and the back cavity (F1 and F3, cf. Apostol et al., 2004). F2 of /y/ is somewhat lower than for /i/.

Speakers can have different strategies for producing the combination of cavity lengths that will lead to the desired acoustic output. For example, they could use more lip protrusion, a more advanced tongue constriction position and a raised larynx, or else less lip protrusion, a more retracted constriction position and a lowered larynx. These two articulatory configurations could both produce the same front and back cavity lengths and a similar acoustic output. The use of different articulatory configurations to produce the same acoustic output has been called *motor equivalence*.

The present study investigates the extent to which speakers will use this particular motor equivalence strategy when compensating for a perturbation of vocal-tract shape. In the first part of the study, participants' speech was perturbed with a palatal prosthesis. There were two kinds of prostheses, one that effectively changed the constriction location of the front-rounded palatal vowel /y/ and one that did not. Our hypothesis was that the speakers with the prosthesis that effectively changed the constriction location would use a motor equivalence strategy (i.e. for example more lip protrusion when the constriction location is fronted with the alveolar palate). Speakers with the other prosthesis should not show this behavior. The speakers' articulator movements were recorded with electromagnetic articulography (EMA).

The second part of the study was designed to investigate whether the two different adaptive behaviors (with different prostheses) could have been governed by a control regime that uses acoustic targets. For this purpose, the adaptation strategies were simulated with the DIVA model. This model of speech production has been shown to be capable of demonstrating a wide range of speech production phenomena; most importantly it has been shown to demonstrate motor equivalence when the vocal-tract of its articulatory synthesizer is perturbed (Guenther et al., 1998). For the current study, the model's vocal-tract shape was adapted to two of our speakers' vocal tracts, with and without prostheses. Then the model was trained to produce /y/ with the unperturbed vocal tract. Afterwards, the adaptation to each type of perturbation was observed. The simulation results were compared with the adaptation data from the two subjects.

2. Experimental data

The first part of the study involved the recording of articulatory (EMA) data of six German speakers, first when they spoke without perturbation, then when they adapted to different prosthesis types.

2.1. Methods

Artificial palates. Our speakers' speech was perturbed by custom-made palatal prostheses. Two types of palatal prostheses were used, one that lowered the palate in the alveolar region and effectively moved the alveolar ridge to a more posterior position ("alveolar prosthesis"), and one that effectively flattened the palatal surface by filling out the palatal vault evenly ("central prosthesis"). All prostheses had a maximum thickness of about one centimeter. Palates were made of dental acrylic and held in place by clasps made from orthodontic wire that fit around the teeth.

Figure 1 shows an example midsagittal contour of each type of prosthesis. The solid thick red lines in each of the panels show the normal palatal contour of the speaker; the blue dashed lines show the perturbed contour. The tongue contour during an unperturbed production of /y/ is shown as thin solid line. The simplest attempt at adaptation when the prosthesis is first inserted would involve a lowering of the tongue. The arrows in figure 1 show the effect of this kind of adaptation on the location of the constriction formed by the tongue. For the central palate (right panel) the constriction location will not change dramatically, so the size of the front cavity will stay the same. For the alveolar palate (left subpanel) however, the constriction will be moved forward towards the location of the artificial alveolar ridge. As a result, the front cavity will become smaller. Speakers with this kind of alveolar prosthesis could then adapt by producing more lip protrusion. Speakers with a central palate should not change lip position very much because the constriction location has not been altered.

Speakers. Six speakers whose first language is German took part in the study, two males (AM1, AM2) and four females (CF1, CF2, CF3, AF1). Three of them, AM1, AM2 and AF1 were provided with a custom-made alveolar prosthesis, the other three, CF1, CF2 and CF3, had a central prosthesis. The speakers were between 25 and 40 years old and spoke Standard German with some regional influence. None of them had a history of speech or hearing problems.

Experimental setup. The articulatory movements of the speakers were recorded with electromagnetic articulography. Sensors were placed midsagittally, three on the tongue, one on the jaw, one on each lip. The front-most tongue sensor was located approximately 1 cm behind the tongue tip, the rear-most sensor opposite the end of the hard palate. Reference sensors for the correction for head movements were placed on the bridge of the nose and on the gingiva above the upper incisors. Data from the upper lip sensor were analyzed as the measure of lip protrusion. For speaker CF1 there was a technical problem with the upper lip sensor which was not noticed until after the recording. Therefore, for this speaker the protrusion of the lower lip sensor was analyzed. Acoustic recordings were made with a microphone connected to a DAT recorder.

Procedure. There were two recordings. In the first session the speakers were recorded without the perturbation (henceforth termed *unperturbed condition*). Then the artificial palate was inserted and the speakers had about 20 minutes to practice speaking with the perturbation. They were then recorded with the prosthesis in place (*perturbed condition*).

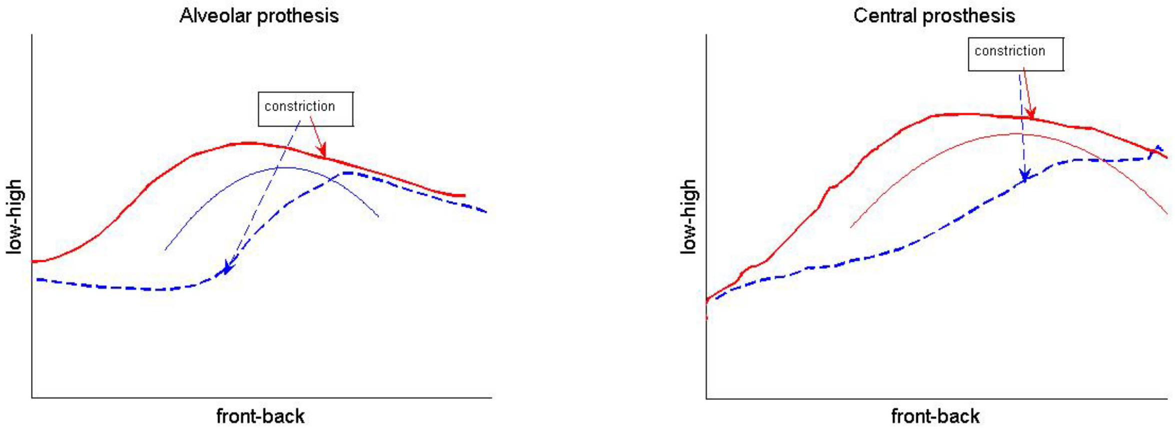

Figure 1: Examples of prosthesis types and their influence on the constriction location during the production of /y/. Front is toward the left. Thick solid red line: natural palatal contour; dashed blue line: prosthesis. Thin line: Tongue contour during the unperturbed production of /y/.

Speech material. The target sound /y/ was embedded in the nonsense word /'ty:ta/, spoken in a carrier phrase: *Ich sah Tüta an* ("I looked at /'ty:ta/."). In order to provide data for building the articulatory model (cf. section 3) further materials (all German lingual sounds) were recorded in CVCV sequences. There were 20 repetitions of each item arranged in random order.

Acoustic analysis. The acoustic signal was downsampled to 24 kHz. The vowel /y/ was segmented based on landmarks (F2 onset to F2 offset) observed in spectrographic display generated from the acoustic signal for each utterance. The first three formants of each produced vowel token were measured manually from the spectrographic display.

2.2 Results

Figure 2 shows the positional measurements of the lip sensor during the different sessions. Data from the speakers with an alveolar palate are shown in the upper row, from the speakers with a central palate, in the lower row. Lower values indicate a more advanced lip position. One can see that all speakers with an alveolar palate have more lip protrusion when the prosthesis is inserted than in their unperturbed speech. The speakers with a central prosthesis show the same lip position in both conditions.

Two-tailed t-tests of unperturbed vs. perturbed conditions were carried out for each of the two prosthesis types. In order to do so the values were z-normalized for each speaker. The results show that for the alveolar prostheses, there was significantly more lip protrusion in the perturbed condition than in the unperturbed condition (p<.001). For the central prosthesis this difference was not significant (p=.134).

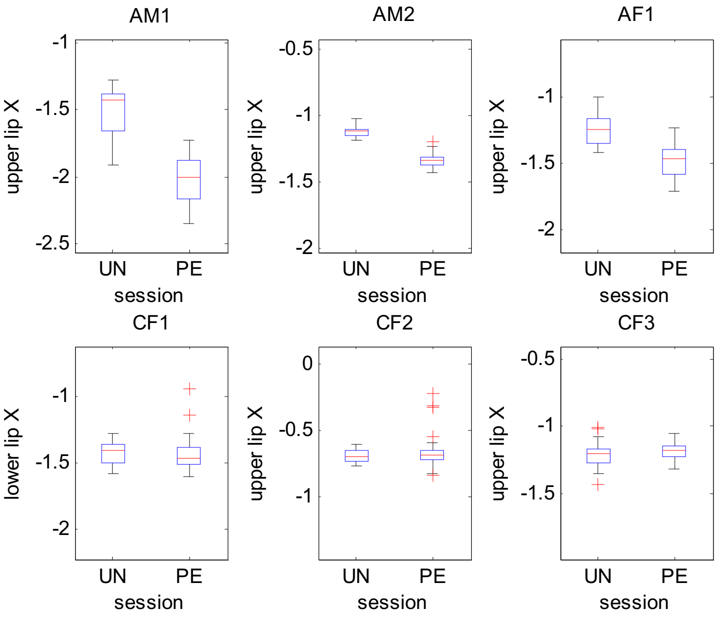

Figure 2: Upper (or lower for speaker CF1) horizontal lip position in cm during unperturbed (UN) and perturbed (PE) speech for speakers with an alveolar palate (upper row) and a central palate (lower row). Higher values denote less lip protrusion.

To summarize, in accord with the hypothesis, the speakers for whom the constriction is presumably fronted by the alveolar palate compensate for the perturbation by using more lip protrusion, thereby lengthening the front cavity.

3. Simulations

In order to further explore the hypothesis that speakers with an alveolar prosthesis are protruding the lips in order to reach a certain acoustic target with their articulators, simulations with the DIVA model of speech production (Guenther et al., 2006) were carried out. This neurocomputational model comprises a controller for a vocal-tract model (Maeda, 1990) and produces vocal tract shapes and acoustic outputs for a given acoustic target. In order to do so, it uses a forward model which is trained during a babbling phase and is capable of predicting the acoustic outcome of a particular articulatory configuration. When the trained model produces a sound or a sound sequence it moves the articulators in directions that yield a match to an acoustic target or sequence of targets.

In the present study the model's vocal tract was adapted to two of our speakers' vocal tracts (AM1 and CF3) in the perturbed and unperturbed conditions. New forward models were learned for these four conditions. Then, the production of /y/ was simulated in the unperturbed and perturbed condition.

3.1 Methods

MRI-recordings. Scans of two of the speakers from the EMA study were performed with a 1.5 Tesla scanner (Philips Achieva X-series), using a neurovascular coil and a T1-weighted, FFE-SENSE sequence. The total acquisition time was 16s. The slice thickness was 2.5 mm (axial slices) and the pixel spacing 0.96x0.96 mm. Subjects were asked to produce either steady state vowels (/a, e, i, o, y, u/) or, for consonants /t, s, ʃ, ç, x, k/, a simple sequence (/aCa/), where the consonantal target position was held during the 16 seconds of image acquisition. Recordings were made at first without the artificial palate, then with the artificial palate in place. The shape of the artificial palate, which could not be seen in the acquired images during most productions, was recorded as well. In order to do so, the tongue was held against the prosthesis so that the prosthesis was completely surrounded by soft tissue, which could be seen on the MRI images.

MRI segmentation. For all recordings, the midsagittal images were aligned with the palatal contour and pharyngeal wall. The midsagittal contour was segmented for all productions. The complete vocal tract shape was segmented for productions of the vowels /a/, /i/ and /u/, with and without the artificial palate in place. These 3D data were needed for conversion of the sagittal outlines to area functions (see below). The artificial palate was segmented as well and combined with the segmentations of the productions with the palate in place.

Articulatory model. An articulatory model was built from the midsagittal contours following a method proposed by Maeda (1990). In order to obtain a sufficient number of midsagittal contours to sufficiently capture the variability of the speaker's productions, data from both EMA and MRI recordings were used. To do so the midsagittal MRI vocal-tract outlines were mapped onto the vocal-tract grid of the Maeda model. Then the positions of the EMA tongue coils for a particular speech sound were mapped onto the segmented midsagittal MRI contours while matching the palatal outline recorded during the EMA recordings with the MRI palatal outline. A linear interpolation was calculated for the tongue contour between the sensors. Then, a complete midsagittal contour was calculated using information from the EMA data if available (in the oral region) and information from the MRI data if no information from EMA was available (in the velar, pharyngeal and laryngeal region). This procedure resulted in 480 tongue contours (20 repetitions per sound * 12 speech sounds * 2 conditions) for each of the two analyzed speakers.

From these tongue contours a jaw movement component was extracted with linear component analysis, taking into account the jaw positions measured from the EMA data. Afterwards, three tongue components (as specified by the Maeda model) were extracted by PCA (tongue position, tongue shape and tongue tip height). Figure 3 shows how varying these four components influences the tongue shape (front is toward the right). The mean tongue position is shown in black, and the tongue contour for maximum and minimum parameter values are shown in green and red, respectively. The jaw component raises and lowers the tongue while retracting it somewhat when the tongue is lowered. The effect of the first tongue component is similar to that of the jaw component. The second tongue component influences the tongue shape (flat vs. bunched). The third tongue component moves the tongue tip. The original Maeda model has components for the configurations of two additional articulators, i.e. lips and larynx, which were left unchanged to reduce the number of articulatory degrees of freedom (DOFs). This was necessitated by the relatively small number of unique vocal tract contours available for articulatory DOF extraction.

The data for the models of the dorsal contour of the vocal tract (alveolar ridge, palate, velar region, pharyngeal wall) were taken from the MRI segmentations and mapped onto the vocal tract grid. There were two dorsal contours for each speaker, one for the unperturbed vocal tract and one for the perturbed one. The model comprising tongue, lips and larynx was combined with one of these two models of the dorsal contour. As a result of this there were two articulatory models for each speaker, one for the unperturbed and one for the perturbed vocal tract, although the articulators (tongue, lips and larynx) were the same for both of these models. Those two articulatory models are shown in figure 6 (black: unperturbed model, red: perturbed model). Thus, there were four articulatory models, representing the perturbed and the unperturbed vocal-tract shape for each of the two subjects. Each of these four models served as the articulatory synthesizer for learning a forward model in simulations with DIVA.

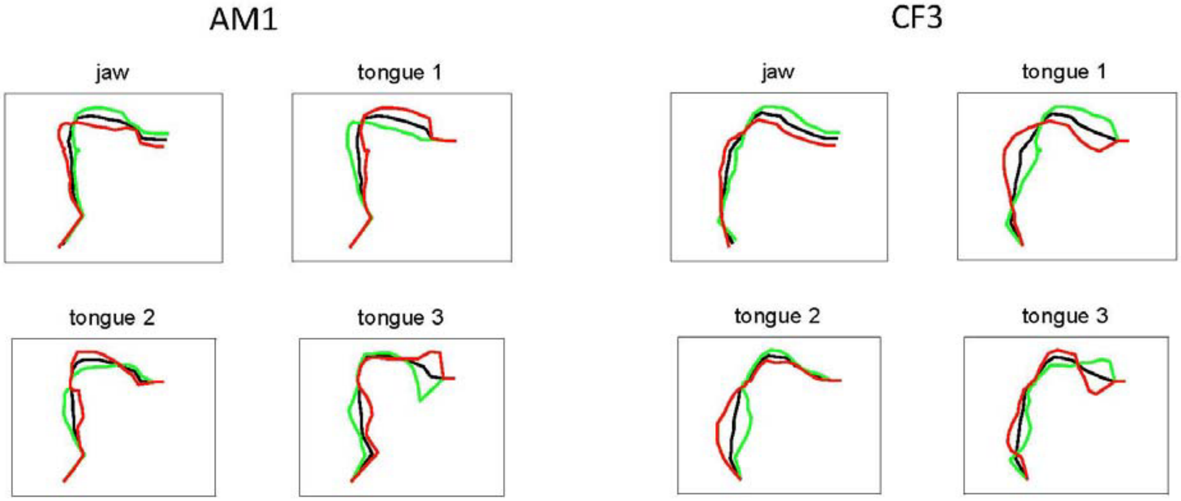

Figure 3. Components of the articulatory model for the speaker with the alveolar palate (left) and the speaker with the central palate (right). Front is right. Black: neutral position, red: parameter value=-3, green: parameter value=+3.

Sagittal-to-area conversion. The sagittal-to-area conversion was performed separately for each speaker and for each condition (unperturbed and perturbed) according to a method proposed by Perrier et al. (1992) while using 3D vocal-tract shape data of /a/, /i/ and /u/. Briefly, this method involves computing the relation between the cross-sectional area *A* and the dorsal-ventral distance *d* using Heinz & Stevens' (1965) formula $A= \alpha * d^{\beta}$, with $\beta=1.5$. The tongue contour and the vocal-tract walls contour in the coronal plane were modeled as parabolic functions of the distance from the mid-sagittal plane. α was then determined for each line of the grid as the ratio $A/d^{1.5}$. It gives a global account of the shape of the cross-section of the vocal tract. Two different α values have been determined for each line of the grid, depending on whether the dorsal-ventral distance is small (below 1cm), or large (above 2cm). For intermediate dorsal-ventral distances an interpolation between the two α values was used.

Forward model. For learning a forward model (during a "babbling phase") that predicts the acoustic output for a given articulatory configuration, a number of syntheses were run, the results of which were then used to train a radial basis function network. The individual steps are shown in figure 4. First, values of the parameters of the articulatory model (jaw, tongue position, tongue shape, tongue tip height, lip protrusion, lip aperture and larynx position) were varied in equal steps ("variation of components", box 2 for the unperturbed model, 2a for the perturbed model). Each of the resulting ~2000 midsagittal vocal-tract shapes was converted to three dimensions, using the sagittal-to-area conversion procedure described above ("Sagittal-to-area conversion", 3 and 3a in figure 4). An acoustic transfer function was calculated using the Maeda synthesizer (Maeda, 1982, 1996, "Synthesis for VT-shapes", 4 and 4a in figure 4). A neural network (the forward model) was then trained to model the functional relation between the articulatory configurations and synthesized outputs ("RBF-network training", 5 and 5a). This was done separately for the perturbed (red in figure 4) and the unperturbed (black) versions of the model, so that there were two forward models, one that would predict the acoustic result for a particular set of values of articulatory parameters for the unperturbed vocal tract and one that would predict the acoustic result for a set of articulatory parameter values of the perturbed vocal tract.

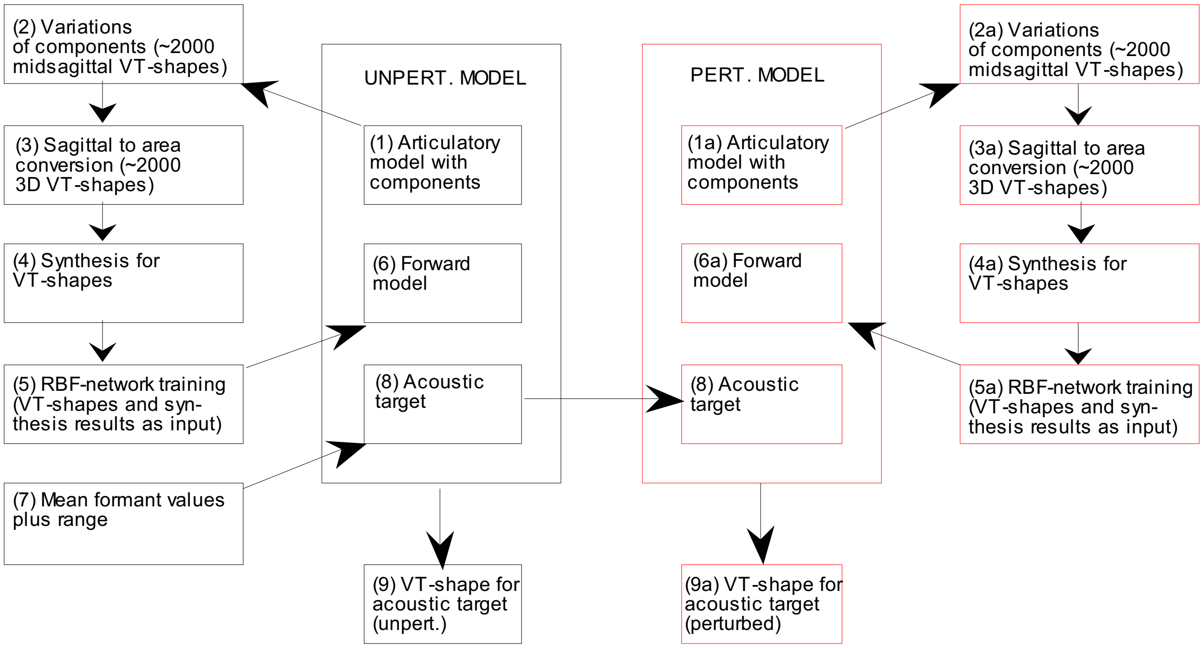
(2) Variations of components (~2000 midsagittal VT-shapes)
UNPERT. MODEL
PERT. MODEL
(2a) Variations of components (~2000 midsagittal VT-shapes)
(3) Sagittal to area conversion (~2000 3D VT-shapes)
(1) Articulatory model with components
(1a) Articulatory model with components
(3a) Sagittal to area conversion (~2000 3D VT-shapes)
(4) Synthesis for VT-shapes
(6) Forward model
(6a) Forward model
(4a) Synthesis for VT-shapes
(5) RBF-network training (VT-shapes and syn-thesis results as input)
(8) Acoustic target
(8) Acoustic target
(5a) RBF-network training (VT-shapes and syn-thesis results as input)
(7) Mean formant values plus range
(9) VT-shape for acoustic target (unpert.)
(9a) VT-shape for acoustic target (perturbed)

Figure 4: Steps during the creation of the unperturbed model (left side, black) and the perturbed model (right side, red).

Acoustic target. An acoustic target for /y/ was estimated for each speaker by calculating mean formant values from the acoustic signals from the 20 unperturbed productions during the EMA recordings. The allowable ranges of the formant values were arbitrarily set to ±40Hz for F1, ±100Hz for F2 and ±200Hz for F3 (box 7).

Simulations. The simulation procedure is also diagrammed in figure 4. The unperturbed model (black), consisting of an articulatory model and a forward model was given an acoustic target (/y/). Then this unperturbed model was trained to produce the vowel /y/ ("vocal tract shape for acoustic target", 9) so that it matched the acoustic target of the speaker (i.e. the formants were within the range defined by the acoustic target). Then, simulations were run in which the perturbed model (consisting of an articulatory model, 1a, and its forward model, 6a) was made to produce an output corresponding to the vowel /y/. To do so, the model adapted and produced a new vocal tract shape for /y/ (9a).

3.2 Results

Figure 5 shows the articulatory configurations produced by the simulations. The left panel shows the results for the model of speaker AM1 (alveolar prosthesis), the right panel, the results for the model of speaker CF3 (central prosthesis).

The unperturbed vocal tract shape is shown with black dashed lines, the perturbed vocal tract, with red solid lines. For the model on the left (alveolar prosthesis) one can see that in the unperturbed condition there is a constriction in the palatal region and some lip protrusion. In the perturbed condition, this model has a more advanced constriction and considerably more lip protrusion. The model on the right (central prosthesis) has a lowered and less bunched tongue in the perturbed condition compared to the unperturbed condition. There is almost no difference in lip protrusion between the unperturbed and perturbed conditions with the central prosthesis. The values of the lip parameters are for the model of AM1 -2.99 in the unperturbed condition and 0.98 in the perturbed condition. For the model of speaker CF3 the difference is marginal (-0.67 in the unperturbed condition and -0.92 in the perturbed condition).

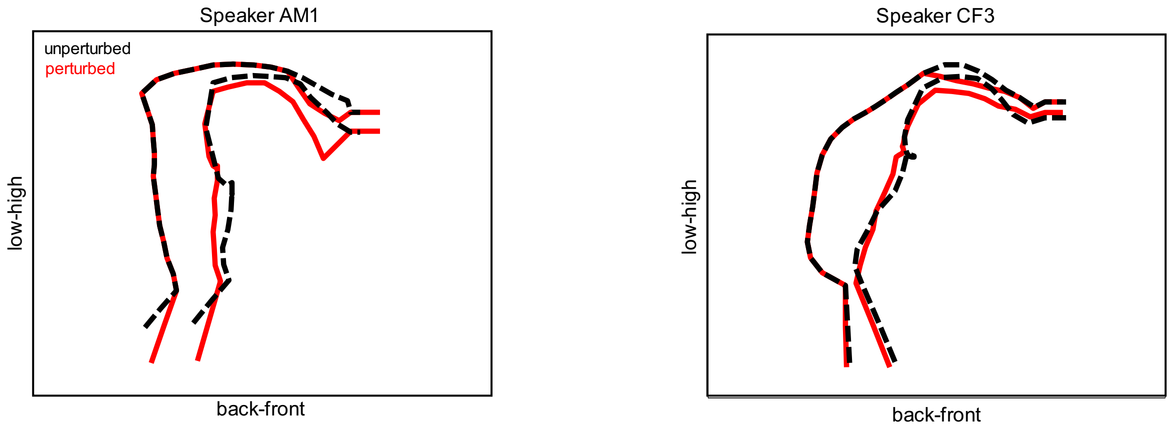

Figure 5: Articulatory configurations produced by the simulations. The model of the speaker with the alveolar prosthesis is shown on the left, the model of the speaker with the central prosthesis is shown on the right. Front is toward the right. The unperturbed production is shown as black dashed line, the perturbed production as red solid line.

Table 1 shows the acoustic results of the simulations and the formant frequencies of the acoustic target. It is evident that all the productions lie within the acoustic target region.

Table 1: Acoustic results

	Model of speaker AM1			Model of speaker CF3		
	F1	**F2**	**F3**	**F1**	**F2**	**F3**
unperturbed	306	1572	1989	320	1893	2416
perturbed	280	1588	2128	283	1901	2413
target	286±40	1650±100	1930±200	320±40	1904±100	2608±200

4. Conclusion

This study has investigated mechanisms of adaptation to a change in the vocal tract shape when speakers produce the vowel /y/. Speakers were provided with one of two kinds of prosthesis. One type of prosthesis (alveolar) was designed to effectively cause a fronting of the tongue constriction for /y/, which was hypothesized to lead to a compensation that involved increased lip protrusion (to maintain the length of the cavity anterior to the constriction). The other type of prosthesis (central) was designed to not change the constriction location; therefore, no compensating change in lip protrusion was expected. Lip position measurements from EMA recordings of two small groups of speakers, one group with each type of prosthesis, supported the hypothesis. The speakers with the alveolar prosthesis demonstrated compensatory lip protrusion, whereas the speakers with the central prosthesis did not.

In the second part of the study, the DIVA model, which employs acoustic targets in controlling articulatory movements, was used in simulations to control an articulatory synthesizer with realistic speaker-specific vocal-tract shapes (unperturbed and perturbed). The simulations show that the observed adaptive behavior can be explained by speakers' attempts to reach a certain acoustic output. Furthermore, the compensatory vocal tract shapes produced by the model show that the tongue constriction location is indeed fronted for the model of the speaker with the alveolar palate and that only this type of palate leads to compensatory lip protrusion.

The results of this study show that speakers are capable of using various articulatory configurations in order to produce a desired acoustic output. The chosen articulatory configuration varies with the overall vocal-tract shape in a way that maintains a stable acoustic output.

Acknowledgements

This research was supported by a grant of the DAAD (German Academic Exchange Service) to Jana Brunner for carrying out postdoctoral research, by grants PO 334/4-1 and HO 3271/1-1 from the Deutsche Forschungsgemeinschaft (B. Pompino-Marschall and P. Hoole, P.I.) and by grant number R01DC00925 from the National Institute on Deafness and Other Communication Disorders, NIH (J. Perkell, P.I.). Thanks to Jörg Dreyer at ZAS Berlin for carrying out EMA recordings, to Laurent Lamalle for carrying out the MRI recordings, to Pierre Badin for help with the setup of the MRI parameters, to Pascal Perrier for a program for the adaptation of α and β values for the sagittal-to-area conversion, to Jonathan Brumberg, Mike Grady, Shanqing Cai and Satrajit Ghosh for advice on working with the DIVA model.

Literature

Apostol, L., Perrier, P. & Bailly, G. (2004). A model of acoustic interspeaker variability based on the concept of formant-cavity affiliation. Journal of the Acoustical Society of America, 115(1): 337-351.

Guenther, F.H., Hampson, M. & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. Psychological Review, 105: 611-633.

Guenther, F.H., Ghosh, S.S., and Tourville, J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language,* 96: 280-301.

Heinz, J.M. & Stevens, K.N. (1965). On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. *Proceedings of the Fifth International Congress of Acoustics*, A44, Liège.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W.J. and Marchal, A. (eds.): Speech Production and Speech Modelling. Dordrecht: Kluwer Academic Publishers: 131-149.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication* 1: 199-229

Maeda, S. (1996). Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer. In: A. Simpson and M. Patzod, Editors, *Arbeitsberichte des Instituts für Phonetik und Digital Sprachverarbeitung der Universität Kiel,* 31: 127–232.

Perrier P., Boë L.J. & Sock R. (1992). Vocal Tract Area Function Estimation From Midsagittal Dimensions With CT Scans and a Vocal Tract Cast: Modeling the Transition With Two Sets of Coefficients. *Journal of Speech and Hearing Research, 35*, 53-67.