ARTICULATORY OPTIMISATION IN PERTURBED VOWEL ARTICULATION

Jana Brunner[1,2,3], *Phil Hoole*[4] *& Pascal Perrier*[2]

[1]Humboldt-Univ. zu Berlin, [2]GIPSA-lab, INP Grenoble, [3]ZAS Berlin, [4]IPSK München
brunner@zas.gwz-berlin.de

ABSTRACT

A two-week perturbation EMA-experiment was carried out with palatal prostheses. Articulatory effort for five speakers was assessed by means of peak acceleration and jerk during the tongue tip gestures from /t/ towards /i, e, o, y, u/. After a period of no change speakers showed an increase in these values. Towards the end of the experiment the values decreased. The results are interpreted as three phases of carrying out changes in the internal model. At first, the complete production system is shifted in relation to the palatal change, afterwards speakers explore different production mechanisms which involves more articulatory effort. This second phase can be seen as a training phase where several articulatory strategies are explored. In the third phase speakers start to select an optimal movement strategy to produce the sounds so that the values decrease.

Keywords: optimisation, motor control, perturbation

1. INTRODUCTION

Several findings from the area of speech motor control and neural processing have led to the assumption that speech motor control involves an internal model in the speaker's brain. In its most general form this model can be seen as some kind of an internal image of the speech production system which serves to predict the acoustic output for a certain motor command input.

Different proposals have been made with regard to the exact structure of the internal model. Most of the proposals include two main mappings, one between motor commands and articulatory configurations and another one between articulatory configurations and sounds. Both these mappings are set up during speech acquisition when the speech production system provides outputs for a given input, namely an articulatory configuration for a given set of motor commands and an acoustic output for a given articulatory configuration.

In this context, speech production planning then would inverse the process by finding a set of motor commands for a given articulatory configuration and an articulatory configuration for a given acoustic output. Since, however, both these mappings are one-to-many this process involves finding a solution to an ill-posed problem.

There are several proposals as to how speakers solve this problem. Most of them can be seen as a kind of optimisation in the sense that the "best" mapping is preferred over the others. They differ in the way in which "best" is defined. One approach (Jordan e.g.[6]) argues that the search for the best mapping involves an optimisation of the articulatory movement via minimisation of the articulatory effort.

In the study presented here we look for evidence for the existence of such an optimisation process. Under perturbation, the speech production system is changed; as a consequence parts of the internal model should also change. Just as during the speech acquisition process, the system is confronted with new mappings of motor commands to articulatory configurations, possibly also with new mappings between articulatory configurations and acoustic outputs. Since the internal model lacks information about the perturbed condition, one can therefore hypothesise that speakers at first will just shift the motor commands in response to the perturbation without changing other movement characteristics, as for example the smoothness of the movement. Since the acoustic output will not in all cases be perfect, one can furthermore hypothesise that in a second phase they will try out other strategies. This will involve increased articulatory effort. In a final phase, they will try to optimise the movement. This should result in a drop of articulatory effort.

Support for this hypothesis of on adaptation process involving several stages comes from previous experimental results for example for the development of accuracy in pointing tasks for children. [3] found that this development is not linear but there are stages where accuracy decreases. These stages are comparable to our second stage where new strategies are tried out, which are not necessarily successful in all cases.

2. METHODS

Palatal prostheses were made for five German native speakers (more speakers are currently being analyzed). For three of them (represented by solid lines in Fig. 1) these prostheses moved the alveolar ridge posteriorily, for the other two (represented by dotted lines) the palatal vault was filled out so that the palate became flatter. Subjects wore the palates for 14 days and were recorded regularly over this period via EMA (Carstens AG 500 for the speaker represented by a black solid line, Carstens AG 100 for the others) in different conditions:

- Day 1:
  - session 1: without artificial palate
  - session 2: with artificial palate, with auditory feedback masking due to white noise over headphones (session missing for one speaker)[1]
  - session 3: with artificial palate with full auditory feedback
- Day 8:
  - session 4: with artificial palate
- Day 15:
  - session 5: with artificial palate[2]

The target sounds /e, i, o, y, u/ were embedded in the nonsense words /'te:ta/, /'ti:ta/, /'to:ta/, /'ty:ta/ and /'tu:ta/ which were produced in the German carrier phrase *Ich sah ... an.* (I looked at ...). The sentences were repeated 20 times per session in randomised order. Three sensor coils were placed on the tongue, one at about 1cm behind the tongue tip, one opposite the border between the hard and the soft palate and a third one in between the two. One further sensor was placed below the lower incisors in order to track jaw movements, one at the upper lip and one at the lower lip. Two coils at the upper incisors and the bridge of the nose served as reference sensors to compensate for head movements.[3] A parallel acoustic recording was carried out on a DAT recorder. The vocalic gestures of the tongue tip (downward movement from the consonant to the vowel) was labeled on the tangential velocity signal using a 20% threshold criterion. The tongue tip sensor was taken because the tongue tip is not involved in fulfilling the task and optimisation should therefore be more evident. Articulatory effort was assessed by means of peak tangential acceleration and tangential jerk ([7]). These measures were chosen because they describe kinematic optimisation and not for example an optimisation in terms of a reduction of the Euclidean distance between two sounds following each other.

Articulatory effort at three points in time was of interest:

- During the unperturbed session. Here one can expect maximally optimised movements since the speaker uses articulatory strategies acquired a long time ago.
- During the perturbation at the point of maximal effort. Maximal effort signals that the speaker tries out new strategies at this stage. Since the effort decreases afterwards this can be seen as the onset of the optimisation.
- During the last perturbed session. Here again one can expect optimised movements since the speakers can be expected to have chosen the most efficient strategy among all the strategies they tried out in phase 2. Possibly optimisation takes longer than our experiment so the optimisation observed here might only be partial.

Repeated measures ANOVAs have been calculated for data split by speaker (SPSS 13, Windows XP). Since the standard error varies for the different sessions, a Greenhouse-Geisser correction of the degrees of freedom has been carried out. Bonferroni posthoc tests for the difference between the unperturbed value and the maximal value, and between the maximal value and the last perturbed value for each of the two parameters were calculated.

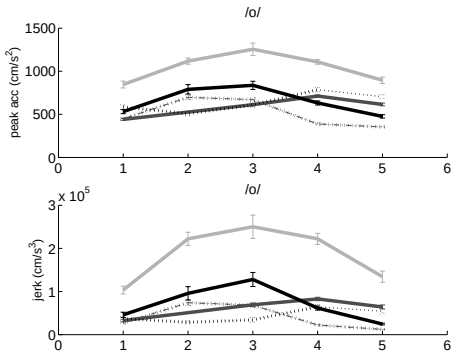
3. RESULTS

The following figure shows the results for the measurements of articulatory effort during the tongue tip gesture towards /o/. The upper panel presents the results for peak acceleration, the lower one the ones for jerk. Different colours and line styles stand for different speakers. The abscissa presents the number of the session as given in the list above. The ordinate presents the value of the parameter. Plots give means and standard error. If the difference between the unperturbed value and the highest value is significant the part of the line from session 1 to the highest value is plotted in bold. If the difference between the highest value and the last session is significant the part of the line from the highest value to the last value is plotted in bold.

For all speakers a significant increase of articulatory effort can be found (difference between session 1 and the session with the maximal values is significant). A significant decrease can be found in all but one case (jerk for the speaker represented by the black dotted line).

The temporal location of the peak points is speaker dependent (e.g. session 3 for the speaker represented by the light grey solid line, but session 4 for two other speakers). Phase 1, where we expected constant values, can, if at all, only be found for the speaker represented by black dotted lines. For the

Figure 1: Optimisation during the gesture towards /o/. Solid lines show results for speakers with alveolar prostheses, dashed lines results for speakers with central prostheses. Different colours and line styles present different speakers. Error bars show standard error.



others it is so short (one or two repetitions) that it cannot be seen in the plot. Phase 2 (with maximal values) can be found for all speakers. Phase 3, the optimisation process (starting at the peak points), can be found for four speakers only, for the fifth the difference between the point with maximal effort (session 4) and the final session is not significant (signaled by the thin line).

For /u/ the results are similar. A significant increase in peak acceleration has been found for all speakers and a significant increase in jerk for four of the five speakers. A significant decrease has been found for all but one speaker for both parameters. Phase 1 can again be found for only for one speaker. For /y/ the results are less clear. Again phase 1 could be found for only one speaker. A significant increase of both, peak acceleration and jerk, has been found for four speakers, a significant decrease for three speakers if measured as peak acceleration and four speakers if measured as jerk. For /i/ a significant increase has been found for only one (peak acceleration) and two (jerk) speakers, a decrease has been found for four speakers if measured as peak acceleration and for three speakers if measured as jerk. For /e/ a significant increase has been found for one speaker only. A decrease has been found for three speakers if measured as peak acceleration and two if measured as jerk.

4. CONCLUSION

To summarise the results, evidence for the existence of phase 1 where the speaker shifts the motor commands but articulatory effort stays constant is difficult to find. This is because this phase is very short, maybe one or two repetitions. For one speaker, however, a rather long phase 1 could be found. Evidence for our hypothesised phase 2, where new strategies are explored and the articulatory effort increases is rather clear. The same is true for phase 3, where the effort decreases because the movements become more and more optimised.

Our results therefore support the hypothesis of the solution of the inverse problem via an optimisation of articulatory effort. As a response to a perturbation speakers compensate in three phases: At first, there is a shift in the system (phase 1). Different articulatory positions are produced but the articulatory effort stays the same. Afterwards, new strategies are explored and the effort increases. Finally, the speakers select the most optimal movement strategies so that the effort decreases.

There is a high dependency on the vowel, the results are clearer for rounded than for unrounded vowels and clearer for back vowels than for front vowels. Furthermore, there is a dependency on the speaker. The speakers represented by the black solid and the dark grey solid line show clearer results than the others. The temporal location of phase 2 (peak point) also depends on the speaker more than on other factors. A possible dependency on the kind of prosthesis can be suspected: The speakers represented by the solid lines (with alveolar prostheses) present clearer results. However, this has to be verified by further experiments with more speakers being carried out at the moment.

For the differences between the vowels several explanations can be found. A reason for the better results for /u/ and /o/ as compared to the front vowels could be that the speakers use the palate as an upper limit ([8]) during the high front vowels so that the possibilities to produce other strategies which possibly involve more articulatory effort are rather limited. This would mean that most speakers do not leave stage 1 during the production of /i/ and /e/.

Another reason for the clearer results for the back vowels is the fact that the acoustics of the two back vowels is not disturbed by the prosthesis since the formants originate in resonances of the back tube ([1]). This means that for the two back vowels the second phase can start immediately after perturbation onset.

A reason for the better results for rounded as compared to unrounded vowels could be that the rounded sounds allow for more variation in tongue position (and therefore more strategies to explore) because acoustic consequences of the variation can be compensated for by different degrees of lip rounding. The investigation of lip protrusion for the vowels is not finished yet, the investigation of /ʃ/ for the same experiment, however, has shown that under pertur-

bation speakers indeed use different degrees of lip rounding for this sound ([2]). One can assume that they use similar strategies for rounded vowels resulting in a greater number of articulatory strategies.

The results could also be interpreted according to a gestural approach. If one assumes that the target words involve coordinative structures the increases in acceleration and jerk could indicate initial difficulty in coordinating C and V gestures which later diminishes with more and more practice. This then results in less articulatory effort.

One could therefore argue that every model, not only the Jordan approach, predicts optimisation. However, it is important to note that the optimisation process here optimises kinematic parameters and not only Euclidean distance towards the target as for example in the approach by Guenther ([4]). This kind of optimisation process is well illustrated by Jordan's model.

It has been remarked that the inverse problem from articulatory configurations to acoustics has been overstated. Evidence has been provided that in normal speech speakers do not use different vocal tract shapes in order to produce the same acoustic output (e.g. [5] and [10] who managed each to recover vocal tract shapes measured on a subject from the acoustic output). However, rather than claiming that the inverse problem has been overstated, our study rather opts for another explanation of this finding. Speakers do not use different articulatory strategies because, once they have found the most efficient strategy, they stay with it.

[1] The influence of feedback masking is not subject of this paper and will therefore not be discussed. During the session speakers could view a sound level display and were instructed to keep their level about equal. Articulatory effort, if it is increased for this session, does therefore not increase as a result of a higher sound level.

[2] A postperturbed session was recorded as well. Since, however, the topic of this paper is optimisation, aftereffects will not be discussed.

[3] [9] report a slight and inconsistent perturbation of speech due to the sensors in EMA recordings resulting in tongue retraction and tongue lowering for some speakers. Since, however, the influence is small and we are furthermore comparing EMA-recordings with each other and not EMA recordings with non-EMA recordings, the effect found between sessions cannot be attributed to the EMA sensors but only to the prosthesis.

5. REFERENCES

[1] 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), 337-350.

[2] Brunner, J., Hoole, P., Perrier, P. & Fuchs, S. 2006. Temporal development of compensation strategies for perturbed palate shapes in German /ʃ/-production. In: Yehia, H.C., Demolin, D. & Laboissière, R. *Proceedings of the 7th International Seminar on Speech Production* 247-254.

[3] Favilla, M. 2006. Reaching Movements in Children: accuracy and reaction time development. *Experimental Brain Research* 169(1), 122-125.

[4] Guenther, F.H. 1995. Speech Sound Acquisition, Coarticulation and Rate Effects in a neural network model of Speech Production. *Psychological Review* 102, 594-621.

[5] Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P. & Saltzman, E. Accurate Recovery of Articulator Positions From Acoustics: New Conclusions Based on Human Data. *Journal of the Acoustical Society of America* 100(3), 1819-1834.

[6] Jordan, M.I. 1989. Indeterminate motor skill learning problems. In: Jeannerod, M. (ed), *Attention and Performance*. Cambridge, MA: MIT Press.

[7] Nelson, W.L. 1983. Physical Principles for Economies of Skilled Movements. *Biological Cybernetics* 46, 135-147.

[8] Stone, M. 1995. How the tongue takes advantage of the palate during speech. In: Bell-Berti, F., Lawrence, J. R. (eds), *Producing Speech: Contemporary Issues for Katherine Safford Harris*. New York: AIP Press, 143-153.

[9] Weismer, G. & Bunton, K. 1999. Influences of pellet markers on speech production behavior: acoustical and perceptual measures. *Journal of the Acoustical Society of America* 105(5), 2882-94.

[10] Yehia, H.C. 1997. *A study on the speech acoustic-to-articulatory mapping using morphological constraints*. PhD dissertation. Graduate School of Engineering of Nagoya University, Nagoya, Japan