Evaluation of an OPG-controlled animated vocal tract model as a biofeedback system

Simon Preuß, Christiane Neuschaefer-Rube, Peter Birkholz

Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital RWTH Aachen
sipreuss@ukaachen.de, cneuschaefer@ukaachen.de, pbirkholz@ukaachen.de

Abstract

We recently proposed an animated vocal tract model controlled by optopalatography (OPG), a technique to track tongue movements inside of the mouth cavity. In this paper, we present a study on the benefit of an improved version of this model in a biofeedback application.To that end, three subjects articulated 16 sustained phonemes, which were recorded using our OPG system. Two of these subjects were then tasked with approximating these 48 target tongue contours and voicing the resulting sounds in two settings, once without visual feedback of their own tongue movements and once with real-time feedback. The results were then evaluated both geometrically (in terms of the distance between the target and approximated contour) and perceptively. We found that adding real-time feedback did reduce the geometric error significantly for one subject, yet this improvement did not lead to a significantly improved perceptive recognition rate. The second subject did not significantly improve neither the geometric nor the perceptive error. However, this pilot study did deliver valuable insights for future improvements on our biofeedback system.

Keywords: OPG, optopalatography, animated vocal tract, biofeedback

1. Introduction

Real-time biofeedback is becoming more and more popular in research, diagnostics, and therapy of speech disorders and pathologies (see, e.g., Engwall 2012 or Richmond and Renals 2012). All currently available techniques to measure speech movements (e.g., cineradiography, electromagnetic articulography, magnetic resonance imaging, ultrasonography), however, have major drawbacks (e.g., cost, precision, complexity) or limitations with respect to temporal and/or spatial resolution. For a biofeedback system to become accepted, it needs to be easily employable with minimum expert knowledge, reliable, precise, and cheap. In our recent and ongoing research (i.e., Birkholz and Neuschaefer-Rube 2011; Birkholz and Neuschaefer-Rube 2012; Birkholz, Dächert, and Neuschaefer-Rube 2012; Preuß, Neuschaefer-Rube, and Birkholz 2013b), we found the technology of optopalatography (OPG) to be a good compromise of these usually mutually exclusive requirements. Using this comparatively simple measurement technique as the data acquisition device, we developed a 2D animation model of the vocal tract that is able to visualize speech movements of a subject in real-time (Preuß, Neuschaefer-Rube, and Birkholz 2013b). In this paper, we present an experiment that was designed to explore the effectiveness of this system for biofeedback. The driving questions were: (1) Can one subject adopt the visually presented midsagittal tongue shape of a sound produced by another subject, (2) does it result in the same sound, and (3) does visual feedback on the tongue movement improve the results?

A similar task is quite common in speech therapy, where the therapist demonstrates a certain tongue movement or position and the patient is asked to imitate it. Without instrumental support, this exercise is quite difficult as neither the patient can exactly see the target nor can the therapist directly assess the precision of the patient's movement. Therefore, only descriptive means and acoustic properties can be applied to evaluate success or failure. Our study was meant to explore whether a visual target and real-time feedback in terms of the 2D tongue shape helps with this difficult imitation task.

2. Optopalatography

Optopalatography (OPG), also called glossometry, is a technique first introduced by Chuang and Wang (1978) and further refined by Fletcher et al. (1989) and Wrench, McIntosh, and Hardcastle (1996). It measures the midsagittal palatolingual distance by means of multiple optical sensors mounted on an artificial palate individually fitted to the subject. Each sensor consists of an infrared LED and a phototransistor. The LEDs are switched on in sequence and send out narrow beams of infrared light which are then diffusely reflected by the tongue below the hard palate. The light intensity at the corresponding phototransistor varies with the distance from the tongue to the hard palate, because less light is reflected to the phototransistor the further the tongue moves away from it. After determining the geometry of the pseudopalate and the optical axes of the sensors, the tongue contour is approximated by linear interpolation between the points marked at the measured distances along the respective sensor axes.This basic principle was further developed by Birkholz and Neuschaefer-Rube (2012) who added another sensor mounted facially of the maxiallary incisors to measure labial aperture.

Our current OPG prototype (see Figure 1) consists of five optical sensors along the midsagittal palate contour and a sixth sensor to measure lip movement: The more light is reflected to that latter phototransistor, the more closed the lips must be. In this study, however, this sensor is not considered, because it can only capture the opening of the lips but not their protrusion (for a more detailed discussion of this problem see Preuß, Neuschaefer-Rube, and Birkholz 2013b). The sensor data are gathered by a microcontroller board and sent to a PC via a serial connection. Our system has a sampling rate of 100 Hz and can therefore acquire data of speech movements in real-time during running speech.

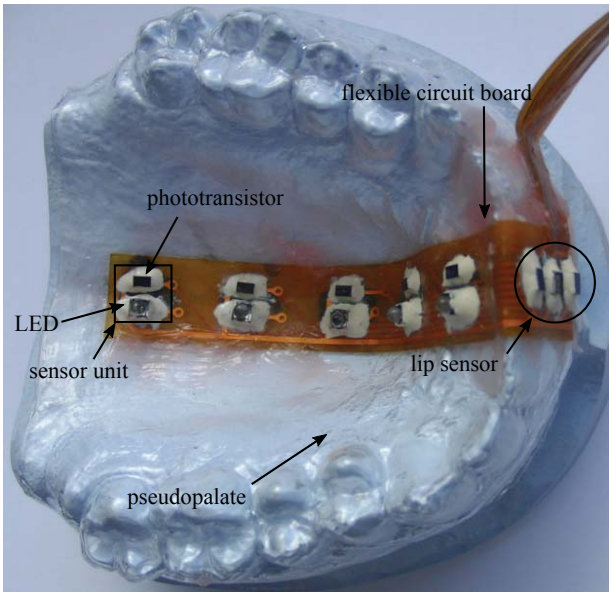

Figure 1: Example of an individually fitted pseudopalate of our current OPG system.

3. The animated vocal tract model

In most OPG systems the tongue contour is constructed by marking a point along the optical axis of each sensor at the respective measured distance and linearly interpolating between these points. In this way, for each sensor unit one point on the tongue surface below the hard palate is obtained. Our system, however, uses a multiple linear regression model to construct the *entire* tongue contour. The basic principle was shown in Preuß, Neuschaefer-Rube, and Birkholz (2013b) for five points and was further developed in Mumtaz et al. (accepted) to obtain 20 points along the entire tongue contour from tongue tip to hyoid, which are then connected by linear interpolation.
For this experiment, we implemented the multiple linear regression model in our 2D animated vocal tract model (see Preuß, Neuschaefer-Rube, and Birkholz 2013b for basic information on the animation model). Also, the ability to load and display previously recorded tongue contours in addition to the currently measured contour was added. The speaker adaptation scheme described in Mumtaz et al. (accepted) was implemented as well to adapt tongue contours recorded with one speaker to the hard palate shape of another speaker.

4. Experimental setup

Three subjects (2 male, 1 female, age 28-35) were outfitted with individually created OPG pseudopalates. The subjects were than asked to voice the eight German vowels /aː, eː, iː, oː, uː, ɛː, øː, yː/ and eight consonants (/f, s, ʃ, ç, ʁ, m, n, l/) for a total of 16 sustained phonemes. During articulation of these sounds, the tongue contours were recorded using our OPG prototype. Each recording consisted of about 2 s of frames of the subjects' tongue contour. To compensate noise and slight variations within the sustained articulation, the average tongue contours were determined by temporal averaging over 500 ms of recorded data. These average contours were used as the 48 target shapes (16 sounds, 3 subjects) for the following task.
The two male subjects (age 28 and 35) were then presented with

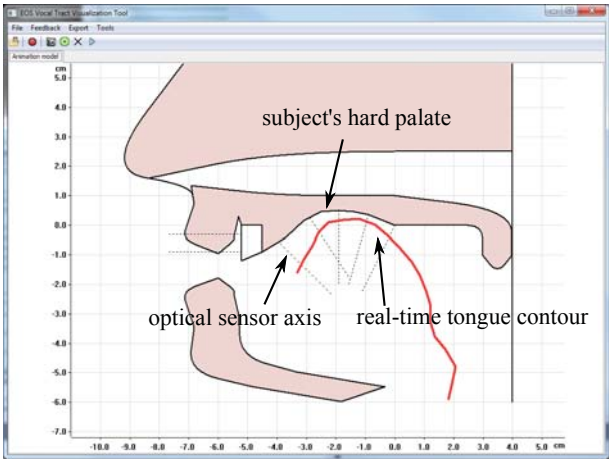

Figure 2: Screenshot of the 2D animation model during articulation of /iː/. The hard palate is loaded individually for each subject. The red tongue contour is updated in real-time. The dotted lines mark the optical axes of the sensors.

these target shapes in randomized order, adapted to their palate geometry by the speaker adaptation scheme mentioned above. The subjects did not know if the presented target tongue contour was from one of their own recordings, or which sound the contour corresponded to. They were tasked with approximating the targets in two settings: once without real-time feedback of their own tongue and once with real-time feedback (see Figure 3). When there was no feedback provided, the subjects had to adjust their tongue position only by intuition. In the setting with feedback, the subjects were given 30 seconds at most to align their tongue contour using the visual aid. Once they deemed the approximation a best fit, the subjects started voicing the resulting sound, and the adopted tongue contour and the acoustic utterance were recorded.

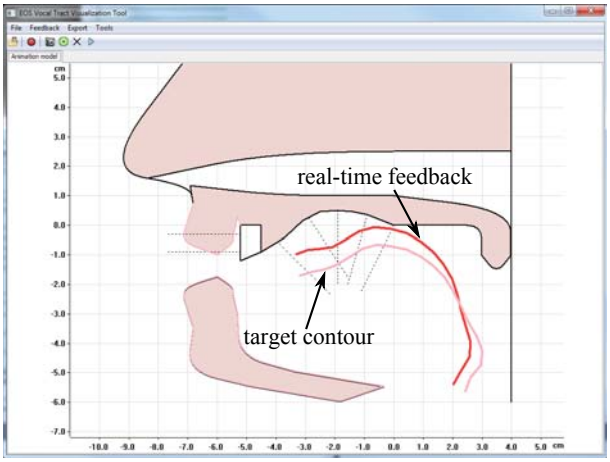

Figure 3: Example of a target contour for /f/ presented in the animation model. The subject's task was to bring its tongue (the red line; only drawn in the condition with feedback) as closely aligned with the light red outlined target contour and then voices the resulting sound. Note that the lips are not shown in their actual shape because the lip sensor currently does not measure the lip protrusion.

5. Evaluation

To quantify the difference between the two test conditions (without and with feedback), we used two error measures: The average geometric distance between the adopted and the target contour and the perceptive recognition rate. The latter was determined by an experienced phonetician who identified the produced phonemes perceptively. The perceived sounds were compared to the actual sounds corresponding to the target contours.

As the geometric error for each of the 20 points, the closest Euclidean distance $d_i (i = 1...20)$ between that point and the target contour was determined. Therefore, the average error as a measure of the goodness of fit was defined as $\sum_{i=0}^{20} d_i$.

6. Results and discussion

Table 1 summarizes the results of the experiment broken down by sound. Even though the recognition rate slightly rises for both subjects when feedback is provided, neither improvement is significant (as determined by Fisher's exact test). The geometric error was also only non-significantly reduced for subject 1 by adding feedback but significantly so for subject 2 (see Figure 4). However, this means that a significant improvement of the geometric error does not necessarily result in an improved overall recognition rate. A possible reason for this is that a single bad fit greatly influences the mean of the distribution. It can be assumed, that a subject achieves the closest approximation when trying to assume a target shape, which was recorded by the same subject. To alleviate the effect of bad fits we therefore now only consider these samples in the following.

	Subject 1				Subject 2			
Sound	without feed-back	reco-gnized	with feed-back	reco-gnized	without feed-back	reco-gnized	with feed-back	reco-gnized
/aː/	0.411	3/3	0.449	3/3	0.948	1/3	0.392	1/3
/eː/	0.268	1/3	0.316	1/3	0.386	1/3	0.223	0/3
/iː/	0.276	2/3	0.276	3/3	0.339	0/3	0.254	1/3
/oː/	0.698	1/3	0.436	1/3	1.027	0/3	0.545	0/3
/uː/	0.31	1/3	0.299	1/3	0.683	0/3	0.55	0/3
/ɛː/	0.274	0/3	0.238	0/3	0.518	1/3	0.289	2/3
/øː/	0.167	0/3	0.173	0/3	0.745	0/3	0.208	1/3
/yː/	0.195	0/3	0.119	1/3	0.195	0/3	0.184	0/3
/f/	0.316	0/3	0.123	0/3	0.464	1/3	0.26	0/3
/s/	0.172	0/3	0.21	1/3	0.574	1/3	0.395	0/3
/ʃ/	0.258	0/3	0.328	0/3	0.486	1/3	0.309	1/3
/ç/	0.303	0/3	0.199	1/3	0.251	0/3	0.239	0/3
/ʁ/	0.553	0/3	0.324	0/3	0.701	0/3	0.449	0/3
/m/	0.236	0/3	0.182	2/3	1.251	1/3	0.588	3/3
/n/	0.162	0/3	0.202	0/3	0.441	0/3	0.26	1/3
/l/	0.199	1/3	0.205	2/3	0.439	0/3	0.267	0/3
Total:		9/48		16/48		7/48		10/48
		$p = 0.1622$				$p = 0.5939$		

Table 1: Average geometric error in cm over all approximated contours and perceptive recognition rate. The p-values were determined with Fisher's exact test (two-tailed).

Table 2 shows the results when the subjects were tasked with approximating their own previously recorded target contours. Adding feedback did not improve the recognition rate for subject 1 at all and could not be found to make a significant difference for subject 2. The error distribution shown in Figure 5 also exhibits no significant improvement of the average geometric error.

The results show that the presented task was much harder than we had anticipated. The overall low recognition rate of no more than 33 % necessitates a much larger number of subjects

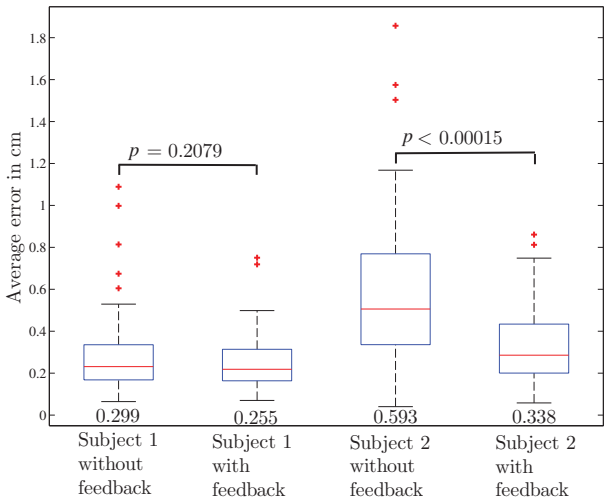

Figure 4: Distribution of the average error in cm over all contours. The numbers below each box are the respective mean of the distribution. Subject 2 achieved a significant improvement in this setting.

and/or items. It was also observed that the adopted tongue contour changed slightly once the subject started voicing the sound (after adjusting it to the target contour without voicing). This aspect is known to be a problem in silent speech interfaces (see, e.g., Denby et al. 2010) and might be investigated further using our feedback system.

The phonetician who evaluated the recordings also noted that the sounds /m/ and /n/ were very hard to discriminate from the audio recordings, further worsening the recognition rate. In a future study, these sounds could be merged to a single class to avoid confusion.

Even though the improvements could not be proven to be significant, we are confident that a larger study will give a more satisfying answer to the question, if this feedback system is useful for this sort of task. For example, one of the subjects showed a very good intuition when approximating some of the target contours even without feedback. More naive subjects without any phonetic knowledge should therefore be considered for the effect of the feedback to manifest more clearly.

In conducting the study we also realized that the information on lip protrusion would greatly benefit the subjects' ability to approximate not only the correct tongue shape but also the correct sound. By further developing the lip feedback, we expect the relationship between geometric error and perceptive recognition rate to become more apparent. The subjects also had no cue, whether to articulate the sounds voiced or unvoiced. By providing this information, the confusion of phonemes could be further avoided. Also, the lateral dimension was completeley missing in the feedback, which made discriminating /l/ and /n/ very difficult for the subjects. By combining OPG with electropalatography (EPG) (e.g., as described in Preuß, Neuschaefer-Rube, and Birkholz 2013a), this information would be available. A future study could also include speech impaired subjects and acoustic targets alongside the visual targets to assess the effect of feedback in a therapeutic context.

Another worthwhile investigation would be to further improve

Sound	Subject 1				Subject 2			
	without feedback	recognized	with feedback	recognized	without feedback	recognized	with feedback	recognized
/aː/	0.086	1	0.263	1	1.084	0	0.251	1
/eː/	0.381	0	0.16	1	0.353	1	0.187	0
/iː/	0.141	1	0.176	1	0.043	0	0.157	0
/oː/	0.492	1	0.179	0	0.727	0	0.382	0
/uː/	0.148	1	0.167	0	1.003	0	0.861	0
/ɛː/	0.176	0	0.228	0	0.722	0	0.515	0
/øː/	0.173	0	0.13	0	0.831	0	0.192	0
/yː/	0.162	0	0.104	0	0.155	0	0.167	0
/f/	0.064	0	0.07	0	0.04	1	0.212	0
/s/	0.082	0	0.169	0	0.575	0	0.146	0
/ʃ/	0.288	0	0.355	0	0.271	0	0.299	1
/ç/	0.292	0	0.224	0	0.125	0	0.298	0
/ʁ/	0.344	0	0.201	0	0.525	0	0.54	0
/m/	0.181	0	0.096	1	0.394	1	0.812	1
/n/	0.109	0	0.207	0	0.307	0	0.14	1
/l/	0.113	1	0.194	1	0.467	0	0.23	0
Total:		5/16		5/16		3/16		5/16
			$p = 1$				$p = 0.6851$	

Table 2: Average geometric error in cm and perceptive recognition rate when approximating own contours. The p-values were determined with Fisher's exact test (two-tailed).

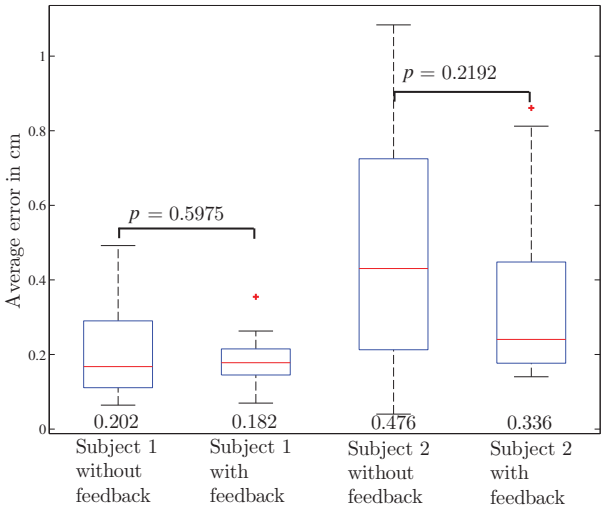

Figure 5: Distribution of the average error in cm when approximating own contours. The numbers below each box are the respective mean of the distribution.

the animation model and the speaker adaptation scheme. So far, these techniques were only evaluated with MRI images of two speakers (see Mumtaz et al. accepted). More data of vocal tract shapes from more speakers are needed to further refine the model and to better adjust the adaptation.

7. Acknowledgements

This work was funded by the German Research Foundation (DFG), grant BI 1639/1-1.

8. References

Birkholz, P., P. Dächert, and C. Neuschaefer-Rube (2012). "Advances in combined electro-optical palatography". In: *Proc. of the Interspeech 2012*. Portland, Oregon, USA.

Birkholz, P. and C. Neuschaefer-Rube (2011). "Combined optical distance sensing and electropalatography to measure articulation". In: *Proc. of the Interspeech 2011*. Florence, Italy, pp. 285–288.

— (2012). "A new artificial palate design for the optical measurement of tongue and lip movements". In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2012*. Ed. by Matthias Wolff. Dresden, Germany: TUDPress, pp. 89–95.

Chuang, C.-K. and W. Wang (1978). "Use of optical distance sensing to track tongue motion". In: *Journal of Speech and Hearing Research* 21, pp. 482–496.

Denby, B., T. Schultz, K. Honda, T. Hueber, JM Gilbert, and JS Brumberg (2010). "Silent speech interfaces". In: *Speech Communication* 52.4, pp. 270–287.

Engwall, O. (2012). "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher". In: *Computer Assisted Language Learning* 25.1, pp. 37–64.

Fletcher, S., M. McCutcheon, S. Smith, and W. Smith (1989). "Glossometric measurements in vowel production and modification". In: *Clinical Linguistics and Phonetics* 3.4, pp. 359–375.

Mumtaz, R., S. Preuß, C. Neuschaefer-Rube, C. Hey, R. Sader, and P. Birkholz (accepted). "Tongue contour reconstruction from optical and electrical palatography". In: *IEEE Signal Processing Letters*. DOI: 10.1109/LSP.2014.2312456.

Preuß, S., C. Neuschaefer-Rube, and P. Birkholz (2013a). "Prospects of EPG and OPG sensor fusion in pursuit of a 3D real-time representation of the oral cavity". In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2013*. Ed. by Petra Wagner. Dresden, Germany: TUDPress, pp. 144–151.

— (2013b). "Real-time control of a 2D animation model of the vocal tract using optopalatography". In: *Proc. of the Interspeech 2013*. Lyon, France, pp. 997–1001.

Richmond, K. and S. Renals (2012). "Ultrax: An Animated Midsagittal Vocal Tract Display for Speech Therapy". In: *Proc. of the Interspeech 2012*. Portland, Oregon, USA, pp. 74–77.

Wrench, A., A. McIntosh, and W. Hardcastle (1996). "Optopalatograph (OPG): A New Apparatus for Speech Production Analysis". In: *4th International Conference on Spoken Language Processing (ICSLP 1996)*. Philadelphia, PA, USA, pp. 1589–1592.