Real-time control of a 2D animation model of the vocal tract using optopalatography

Simon Preuß, Christiane Neuschaefer-Rube, Peter Birkholz

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

Abstract

This paper presents an animated 2D articulation model of the tongue and the lips for biofeedback applications. The model is controlled by real-time optopalatographic measurements of the positions of the upper lip and the tongue in the anterior oral cavity. The measurement system is an improvement on a previous prototype with increased spatial resolution and an enhanced close-range behavior. The posterior part of the tongue was added to the model by linear prediction. The prediction coefficients were determined and evaluated using a corpus of vocal tract traces of 25 sustained phonemes. The model represents the tongue motion and the lip opening physiologically plausible during articulation in real-time.
Index Terms: optopalatography, glossometry, linear prediction, animated vocal tract model, biofeedback

1. Introduction

The therapy of speech and communication disorders is currently heavily reliant on the auditory skills of the therapist and the subject. Visual feedback of articulation is only sparsely available even though knowledge of the position and movement of the articulators is crucial to effectively instruct a subject, as shown by several studies (e.g., [1, 2]). A visual representation of the vocal tract based on real-time measurements would also greatly benefit a wide scope of non-therapeutic applications, such as silent speech interfaces, speech prostheses or articulatory speech synthesis.

In order to visualize a subject's vocal tract, a system needs to extract information about the articulators either by a process called acoustic-to-articulatory inversion or by directly measuring the articulators in question. Acoustic-to-articulatory inversion attempts to infer the varying shape of the vocal tract from the acoustic speech signal. However, the non-uniqueness of the inversion poses a major challenge as a given acoustic signal can be produced by multiple different vocal tract shapes. For that reason, several attempts have been made to extend the problem to audiovisual-to-articulatory inversion by correlating facial movements with articulatory properties [3, 4, 5]. Another approach was to incorporate phonetic, phonological and dynamical knowledge and constraints into the inversion [6, 7, 8]. Still, while some progress has been made in this field, the non-uniqueness remains an obstacle.

However, to measure the articulators directly in a practical way is also a non-trivial task. The current state of the art in instrumental speech therapy and articulation analysis mainly employs rather complex and intricate techniques (e.g., ultrasound, electromagnetic articulography (EMA), or X-ray microbeam) [9]. Although these procedures can accurately measure and visualize the motion and position of the tongue and the lips, all of them rely on expensive equipment and specialized personnel. Their application is also quite strenuous for the subject and interferes with the freedom of movement of the articulators. For example, Steiner and Ouni [10] developed an animated EMA controlled tongue model. This technique is however prone to errors as, according to [10], EMA position and orientation data are not very reliable and the fragile coils are ill-suited for a monitoring or feedback application. To remedy these drawbacks, Richmond and Renals [11] developed an animated vocal tract model that selects vocal tract shapes from a corpus of predefined shapes based on a few control parameters obtained by ultrasound measurements during articulation. Their work is however still in the early stages.

An alternative, much simpler approach to measure the tongue contour in the anterior oral cavity was introduced in 1978 by Chuang and Wang [12]. Their proposed device comprised four optical sensors mounted on the midsagittal line of an artificial plate (pseudopalate), which was molded to fit the subject's palate. These sensors determined the distance of the tongue from the pseudopalate by measuring the light intensity reflected by the tongue surface. The principle was refined and dubbed glossometry by Fletcher et al. [13] in the 1980s and further developed by Wrench et al. [14, 15] into the optopalatograph, but never went on to see any widespread use, with the expensive manufacturing process of the handmade palates being a major factor. Though it holds great potential, optopalatography (OPG) has so far only been able to measure the tongue contour in the anterior oral cavity, ignoring the posterior part of the tongue and the lips.

In this study we present a prototype system consisting of an optopalatographic measurement device and a visualization model. The measurement device improves on [16], who added a forward-directed sensor in front of the upper incisor to measure the aperture of the lips. The visualization model extends the information gained by OPG through linear prediction to estimate additional points on the tongue surface in the posterior region based on the measured data. The measured and predicted data control the lips and tongue in an animated real-time 2D model of the vocal tract, whose dimensions were obtained from physiological magnetic resonance imaging (MRI) vocal tract traces.

2. Optopalatography

The proposed OPG system uses five optical sensors, mounted on a flexible circuit board that is affixed to a pseudopalate, to measure the distance from the palate to the tongue, and one sensor unit placed on the labial side of the upper incisors to measure the lip aperture. It is minimally invasive and interferes with

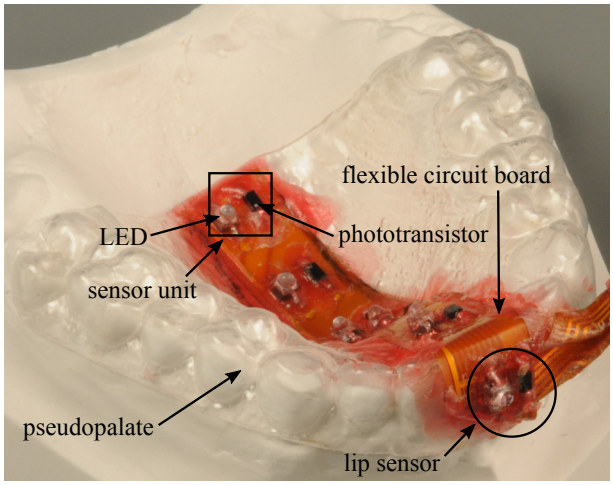

Figure 1: Assembled pseudopalate fixed on a plaster model of the subject's maxilla.

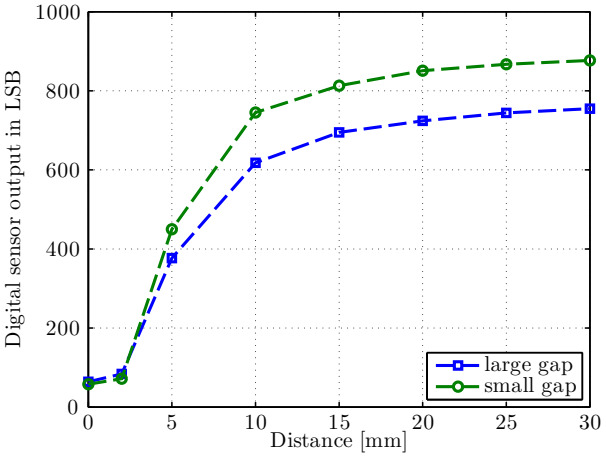

Figure 2: Distance sensing functions for two sensor units (LSB: least significant bit) with different layouts: The green line with circle markers represents a unit with a smaller gap of 3.2 mm between the LED and the phototransistor and the blue line with square markers is a unit with a larger gap of 3.8 mm. For details on the calibration process see [16].

the subject's articulation and general freedom of movement to a much lesser degree than established procedures. Its low technological requirements enable data acquisition in real-time and its simple manufacturing process and handling allow a cheap, portable and user-friendly implementation, thus facilitating its use in clinical, research or even home applications.

Each sensor consists of an infrared light-emitting diode (LED), which emits a light-beam on the tongue or lip surface, and a phototransistor that collects the reflected light and transforms it into a proportional current. The registered light intensity is then related to the distance of the reflective tongue surface to the sensor and the degree of opening of the lips, respectively (i.e., the further away the tongue or the more open the lips, the less light is reflected to the phototransistor). For additional details on choice and calibration of the optical sensors see [16, 17, 18].

Before the sensor units are mounted on the flexible circuit board, the bottom of the clear casings of the LEDs is coated with biocompatible black nail polish to avoid an optical bridge through the circuit board material between an LED and a phototransistor. The sensor board is then glued to an artificial palate that is individually molded to fit the subject's hard palate. This pseudopalate's base layer is merely 0.5 mm thick and therefore interferes only marginally with the subject's articulation. The LED and phototransistor contacts, as well as the edges of the components and the circuit board, are sealed with pattern resin, a modeling resin used in dentistry and suitable for application in the oral cavity. The sealing prevents saliva from running under the components and ensures that the subject cannot come into contact with electrically charged elements. All wires necessary to read and control the sensors are bundled and exit through the mouth opening, wrapped in a thin silicon foil for additional wearing comfort. The final system is shown in Fig. 1.

It was also examined how the gap between the phototransistor and the LED of a sensor unit influences the measured values at close range by comparing two sensor units, one designed with a distance of 3.8 mm (center to center) and one with a distance of 3.2 mm between the components. Fig. 2 shows the distance sensing functions for both setups. It is evident that reducing the gap between the LED and the phototransistor leads to no significant improvement in terms of distance resolution, especially at close range. Because of the easier handling when sealing the sensors, it was advisable to use the layout with the larger gap in all subsequent measurements. Both layouts are superior to the previous prototype in [16] due to the black coating of the LED casings and the addition of a fifth sensor along the midsagittal line for an increased spatial resolution.

3. Animation model

The animation model provides a two-dimensional, real-time representation of the vocal tract using the data acquired by OPG to animate the tongue and the lips. The tongue contour is extended beyond the measured points by linear prediction. The geometry of the hard palate was immediately available by measuring the pseudopalate. The model allows different pseudopalate description files to be loaded that contain up to 64 outline points as well as the positions of the optical sensor units and their respective optical axes. The remaining geometric dimensions of lips, velum, nose, soft palate and posterior pharyngeal wall were determined by measuring their counterparts in a corpus of vocal tract traces that also served as training data for the linear prediction coefficients.

3.1. Data

The corpus of vocal tract traces was obtained in [19] from volumetric MRI data acquired for [20] that consisted of sustained articulations of German vowels and consonants. Of this corpus we used the traces of the long vowels /aː/, /eː/, /iː/, /oː/, /uː/, /ɛː/, /øː/, and /yː/, the short vowels /ɪ/, /ɛ/, /a/, /ɔ/, /ʊ/, /ʏ/, /œ/, /ə/, and /ɐ/, and the consonants /f/, /l/, /m/, /n/, /s/, /ʃ/, /ç/ and /x/ (25 sounds in total) as training data for the linear prediction of five points on the posterior part of the tongue surface. The idea was to predict the $x-$ and $y-$components of these points $\hat{P}_j(x, y)$ $(j = 1, ..., 5)$ by linearly combining the measured palato-lingual distances d_i $(i = 1, ..., 5)$ according to Eqs. (1) and (2), respectively.

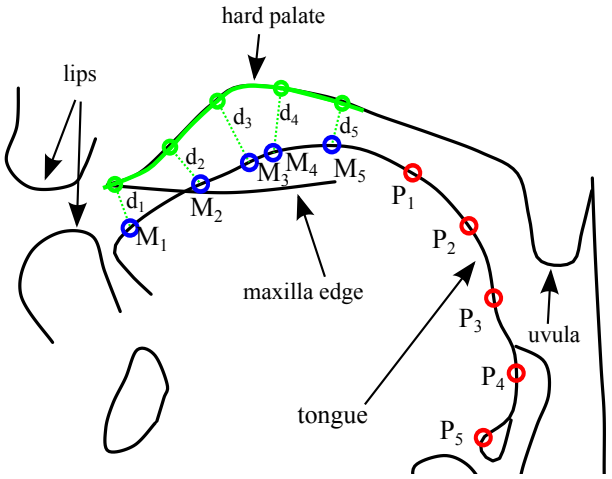

Figure 3: Annotated trace of the vowel /ɛː/. Pseudopalate, estimated sensor positions and optical axes are green, known distances d_i and points M_i are blue, prediction training points P_j are red.

$$\hat{P}_{j,x} = a_{j,0} + \sum_{i=1}^{5} a_{j,i} \cdot d_i \quad (1)$$

$$\hat{P}_{j,y} = b_{j,0} + \sum_{i=1}^{5} b_{j,i} \cdot d_i \quad (2)$$

In order to determine the coefficients $a_{j,0}, ..., a_{j,5}$ and $b_{j,0}, ..., b_{j,5}$, a "virtual pseudopalate" was inserted into the vocal tract traces, i.e., five reference points corresponding to the optical sensor positions on an actual pseudopalate were marked in the traces of the analyzed phonemes along the hard palate. The most posterior sensor unit designated the origin of the reference coordinate system and the edge of the maxilla was aligned parallel to the reference x-axis. From these points, OPG measurements were mimicked by determining the optical axes of the sensors and their intersections with the tongue surface, yielding five distances d_i and hence five points M_i on the tongue surface for each phoneme. Five additional points $P_j(x, y)$ were equidistantly marked along the tongue contour between the most velar point M_5 and the tongue root (see Fig. 3). In this way five overdetermined sets of 25 linear equations were obtained for each component that were then solved for $a_{j,0}, ..., a_{j,5}$ and $b_{j,0}, ..., b_{j,5}$, respectively, using a standard MATLAB least square error (LSE) algorithm.

3.2. Animation control

The system acquires OPG data with a frame rate of 100 Hz. For the palate sensors, the measured light intensity is mapped to a palato-lingual distance by linearly interpolating between predetermined calibration values, which are stored in palate description files. The five points M_i on the anterior tongue surface are obtained by marking a point on the optical axis of each sensor at the measured distance d_i, respectively. The angles of the optical axes are also stored in the palate description files. The light intensity registered by the lip sensor is translated into a numerical opening factor between 0 and 100, where 0 describes the lip position of the sound /m/ (i.e., completely closed) and 100 corresponds to the lip position of the sound /iː/ (i.e., open). The lip

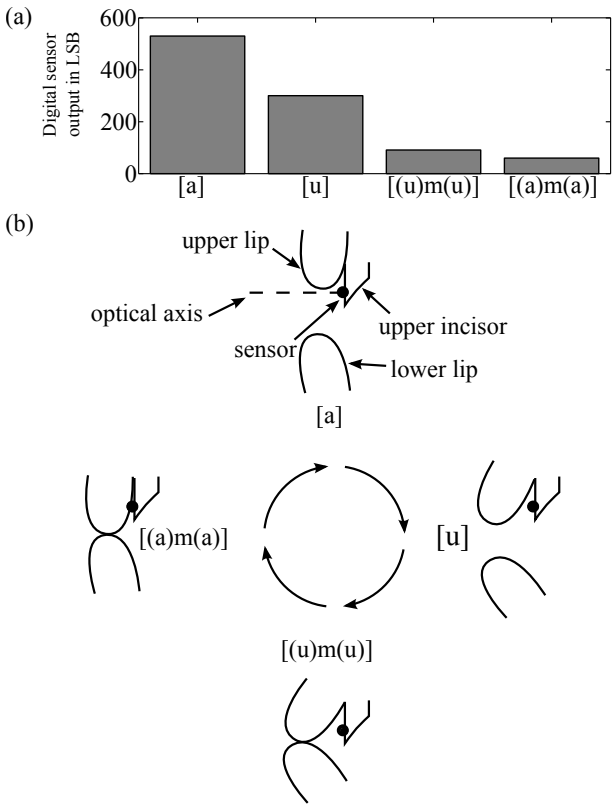

Figure 4: Lip sensor output interpretation: (a) Lip sensor output for different sounds (LSB: least significant bit). (b) Cycle of possible degrees of opening and protrusion when using only one lip sensor.

sensor only measures the lip opening while in fact there are two degrees of freedom (protrusion and opening) in lip movement. It is however not possible to capture both properties with a single sensor. Fig. 4 (a) shows the sensor output for varying degrees of opening and protrusion. If the lips are completely open and not protruded (e.g., for the sound /aː/), very little light is reflected as the lip does not occlude the sensor. In case of open and protruded lips (e.g., during articulation of /uː/), the sensor output is lower, as a little more light is reflected from the lip that is now covering the sensor at a small distance in front of it. If the lips are closed while the protrusion remains the same (e.g., for /m/ in an /u/ context), much more light is reflected as the lips completely cover the sensor and no light can escape through the mouth opening. For closed but not protruded lips (e.g., for /m/ in an /a/ context), the registered light intensity peaks as the sensor is completely occluded at a minimum distance. This means that when moving from one sound to another, only the path shown in Fig. 4 (b) can be assumed in the model while in fact the actual gesture might go from open and not protruded to closed and not protruded directly. Therefore the transitions between the described extremes are non-unique using a single lip sensor and cannot be reconstructed accurately and the visualization must be limited to the opening.

The third part of the animation model, the posterior part of the tongue contour, is calculated by linearly combining the measured palato-lingual distances as described in Sec. 3.1. After all points are calculated, they are drawn on the screen and linearly interpolated to form the tongue contour. The lips are drawn at a closed resting position for an opening factor of 0 and the upper

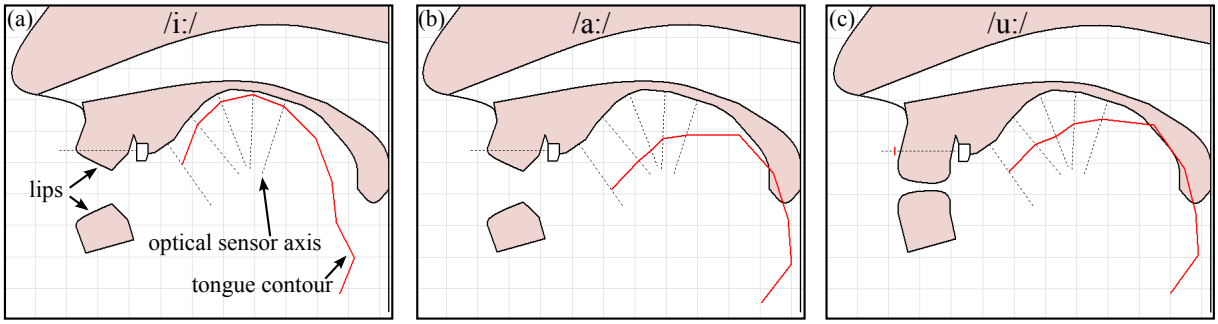

Figure 5: Still frames of the animation model taken from a real-time measurement using the same speaker with the tongue contour painted in red: (a) Front vowel /i:/. (b) Low back vowel /a:/. (c) High back vowel /u:/.

lip is raised in proportion to the opening factor while the lower lip is drawn as the upper lip, mirrored at the line of lip closure at the resting position. The lip animation therefore assumes symmetrical movement of lower and upper lip, which is acceptable for visualization purposes.

4. Evaluation

To find a measure of confidence in the prediction of the tongue contour points, the root mean square (RMS) deviation of the predicted points $\hat{P}_j$ from the corresponding training points P_j and the coefficient of the correlation between $\hat{P}_j$ and P_j over all 25 training samples were calculated and are shown in Fig. 6. The correlation is generally significant ($p < 0.01$).

Fig. 5 shows three examples of the animation model state for the sustained vowels /i:/, /a:/ and /u:/. The still frames were taken from real-time measurements of the respective sustained sound using the same speaker and palate configuration (see also supplemental files for example utterances). The anterior region, which is reconstructed based on the OPG measurements, shows a much more elevated tongue for /i:/ than for the other sounds. The lip opening is large for /i:/ and /a:/, and much smaller for /u:/, which is physiologically correct. The predicted tongue contour points of the back vowels /a:/ and /u:/ are realistically further posterior than in the case of the front vowel /i:/.

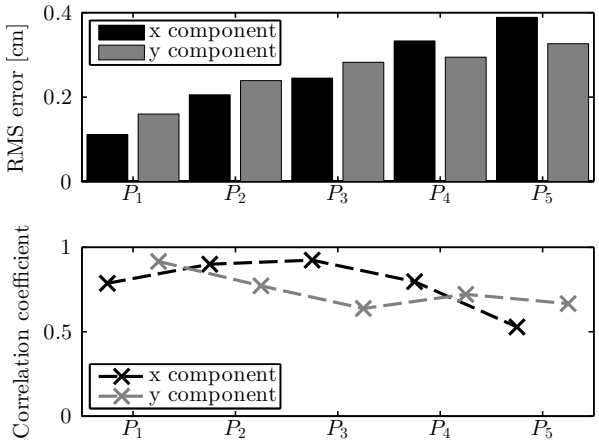

Figure 6: Root-mean-square error in cm and correlation coefficient of the x (black) and y (gray) components of each predicted point with respect to the corresponding training point. The correlation coefficients are in relation to the training data points; $p < 0.01$.

5. Discussion and conclusions

The evaluation of the animation model has shown that the acquired tongue contours are in line with physiological tongue movements and positions during the articulation of the observed sounds and thus can be used for qualitative examinations of a subject's midsagittal vocal tract. It has been shown that simple linear prediction is highly effective for predicting the posterior tongue shape from the anterior contour. To thoroughly evaluate the performance of the system however, a ground truth needs to be established to which the determined tongue contours can be compared. To that end, traces of the vocal tract of the sounds that were used to control the animation model articulated by the same speaker are required and need to be acquired (e.g., by MRI) in a future study. Once this ground truth is known, it can also be quantitatively examined how well the prediction of the tongue contour can be generalized from one speaker to another.

The current system uses a single lip sensor while the lips actually have two degrees of freedom: opening and protrusion. It was shown that a single sensor is not sufficient to allow for smooth and physiologically accurate transitions, and a future study should examine if a second lip sensor can solve this indeterminacy. The movement of another articulator, the velum, is not represented at all in the current system. In a future study, we will therefore examine if the velum motion has a measurable effect on the intensity of transilluminating infrared light emitted from an LED at the nostrils. In that case the variation of light intensity could be mapped to a degree of opening and yield control data to animate the velum in the model.

In addition to this, the model could be generalized into a three-dimensional representation of the vocal tract by adding another technique based on a pseudopalate, namely electropalatography, as conceptually described in [21]. As a long-term goal, the information gained by a three-dimensional model could be used to control a silent speech interface or to excite an articulatory speech synthesis model by directly providing the input for the articulatory parameters.

6. Acknowledgements

This work was funded by the German Research Foundation (DFG), grant BI 1639/1-1. We wish to thank Antigone Kirchharz for her help in the production of the plaster model of the palate.

7. References

[1] D. M. Ruscello, "Visual feedback in treatment of residual phonological disorders," *Journal of Communication Disorders*, vol. 28, no. 4, pp. 279 – 302, 1995.

[2] P. Badin, A. B. Youssef, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *L2SW, Workshop on" Second Language Studies: Acquisition, Learning, Education and Technology*, 2010, pp. P1–10.

[3] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.

[4] J. Jiang, A. Alwan, L. E. Bernstein, P. Keating *et al.*, "On the correlation between facial movements, tongue movements and speech acoustics," in *The 6th International Conference on Spoken Language Processing*, 2000, pp. 42–45.

[5] H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion," *Speech Communication*, vol. 51, no. 3, pp. 195–209, 2009.

[6] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 237–240.

[7] S. Dusan, "Methods for integrating phonetic and phonological knowledge in speech inversion," *Proceedings of the International Conferense on Speech, Signal and Image Processing, Malta*, 2001.

[8] B. Potard, Y. Laprie, and S. Ouni, "Incorporation of phonetic constraints in acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 123, p. 2310, 2008.

[9] M. M. Earnest and L. Max, "En route to the three-dimensional registration and analysis of speech movements: instrumental techniques for the study of articulatory kinematics," *Contemporary Issues in Communication Science and Disorders*, vol. 30, pp. 2–25, 2003.

[10] I. Steiner and S. Ouni, "Progress in animation of an EMA-controlled tongue model for acoustic-visual speech synthesis," in *Elektronische Sprachsignalverarbeitung*. TUDpress, 2011, pp. 245–252.

[11] K. Richmond and S. Renals, "Ultrax: An animated midsagittal vocal tract display for speech therapy," in *Proc. Interspeech*, Portland, Oregon, USA, 2012.

[12] C.-K. Chuang and W. S. Wang, "Use of optical distance sensing to track tongue motion," *Journal of Speech and Hearing Research*, vol. 21, pp. 482–496, 1978.

[13] S. G. Fletcher, M. J. McCutcheon, S. C. Smith, and W. H. Smith, "Glossometric measurements in vowel production and modification," *Clinical Linguistics and Phonetics*, vol. 3, no. 4, pp. 359–375, 1989.

[14] A. A. Wrench, A. D. McIntosh, and W. J. Hardcastle, "Optopalatograph (opg): A new apparatus for speech production analysis," in *4th International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA, USA, 1996, pp. 1589–1592.

[15] A. A. Wrench, A. D. McIntosh, and W. J. Hardcastle, "Optopalatograph: Development of a device for measuring tongue movement in 3d," in *EUROSPEECH '97*, Rhodes, Greece, 1997, pp. 1055–1058.

[16] P. Birkholz and C. Neuschaefer-Rube, "A new artificial palate design for the optical measurement of tongue and lip movements," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2012*, M. Wolff, Ed. Dresden, Germany: TUDPress, 2012, pp. 89–95.

[17] P. Birkholz and C. Neuschaefer-Rube, "Combined optical distance sensing and electropalatography to measure articulation," in *Interspeech 2011*, Florence, Italy, 2011, pp. 285–288.

[18] P. Birkholz, P. Dächert, and C. Neuschaefer-Rube, "Advances in combined electro-optical palatography," in *Proc. of Interspeech 2012*, Portland, Oregon, USA, 2012.

[19] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.

[20] B. J. Kröger, R. Winkler, C. Mooshammer, and B. Pompino-Marschall, "Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results," in *5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Bavaria, 2000, pp. 333–336.

[21] S. Preuß, P. Birkholz, and C. Neuschaefer-Rube, "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*, P. Wagner, Ed. TUDPress, Dresden, 2013, pp. 144–151.