Tongue Contour Reconstruction from Optical and Electrical Palatography

Rizwan Mumtaz, Simon Preuß, Christiane Neuschaefer-Rube, Christiane Hey, Robert Sader, and Peter Birkholz, *Member, IEEE*

***Abstract*—Tongue shape reconstruction based on safe and convenient measurement techniques is of great interest for speech research and speech therapy. Two potentially useful and related measurement techniques for this purpose are electropalatography (EPG) and optopalatography (OPG). While EPG measures the time-varying contact pattern between the hard palate and the tongue, OPG measures distances between the two. Here, we examined the potential of EPG, OPG, and their combination for predicting the whole tongue contour using a multiple linear regression model. The model was trained and tested with tongue shapes and virtual sensor data obtained from Magnetic Resonance Images of sustained articulations of two speakers. When the model was trained and tested with the same speaker, the error of tongue contour reconstruction was significantly lower for predictions based on OPG data than for predictions based on EPG data. When the model was trained with one speaker and tested with the other, the error pattern was less consistent and the overall error was higher. Hence, especially OPG is well suited for tongue contour prediction, but an adaptation method is needed to transfer the model to a new speaker.**

***Index Terms*—Electropalatography, magnetic resonance imaging, multiple linear regression, optopalatography.**

I. Introduction

TONGUE shape reconstruction from sensor data is of great interest for applications in speech research and speech therapy. It can be used, for example, to provide visual articulatory feedback for patients with speech disorders (e.g., [1], [2]) or to drive articulatory models for speech synthesis (e.g., [3]). There are many methods to obtain data related to the shape of the tongue, e.g., ultrasonography (US), electromagnetic articulography (EMA), cineradiography, X-ray microbeam, magnetic resonance imaging (MRI), electropalatography (EPG), and optopalatography (OPG) [4]. Each method has specific limitations with respect to temporal and spatial resolution of the data, safety, usability, and cost. Cineradiography is rarely used to study speech production anymore because of the harmful radiation. While US, EMA, X-ray microbeam, EPG, and OPG have a high temporal resolution (e.g., 100 Hz), the spatial detail they provide is limited. For example, EMA provides the position of only a few flesh points on the tongue surface, and EPG provides only the pattern of contact between the tongue and the hard palate. MRI, on the other hand, can provide very detailed images of the tongue, but needs long acquisition times.

However, many applications would benefit from articulatory data with both a high temporal resolution and spatial detail. To achieve this, predictive models can be created to map low-dimensional sensor data with a high temporal resolution to spatially detailed tongue shapes. Along this line, multiple studies analyzed the prediction of the tongue contour as obtained by US or X-ray measurements from the position of a few points on the tongue surface as measured by EMA [5]–[9]. However, X-ray is harmful, and US does often not capture the *complete* tongue contour due to a limited measurement angle of the probe. Furthermore, EMA measurements are rather inconvenient for the subject and hence impractical as input to the predictive model in clinical use, for example.

In the present study, we analyzed the potential of EPG and OPG for tongue contour reconstruction, building on the previous pilot study [10]. These techniques are safe, convenient, and rather cheap, and EPG is established in many speech therapy clinics [11]. Both methods are similar in that the subject has to wear an artificial palate fitted to the shape of his hard palate. This artificial palate is equipped with a grid of electrodes to detect contact between the tongue and the palate for EPG [11], and with multiple optical sensors to measure distances from the palate to the tongue surface for OPG [12]–[15]. Figs. 1(a) and (b) show examples of an EPG and an OPG palate, respectively. The EPG palate has 62 contact sensors distributed on its surface, and the OPG palate contains five distance sensors directed towards the tongue. Fig. 2 illustrates the kind of data provided by EPG and OPG, i.e., a binary pattern of palatolingual contact and five measures of distance from the palate to the tongue surface. While the distance patterns directly translate to individual points on the anterior tongue, the contact patterns provide only indirect information about the tongue shape. However, it is plausible to assume that EPG patterns and tongue shapes are strongly related. The aim of this study was to explore to what extent a multiple linear regression model can predict the *whole* con-

Manuscript received December 20, 2013; revised March 14, 2014; accepted March 15, 2014. Date of publication March 21, 2014; date of current version March 27, 2014. This work was supported by the BMBF under Grant 13EZ1125A and by the German Research Foundation under Grant BI 1639/1-1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Arrate Munoz-Barrutia.

R. Mumtaz, S. Preuß, C. Neuschaefer-Rube, and P. Birkholz are with the Department for Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen, RWTH Aachen University, 52074 Aachen, Germany (e-mail: rmumtaz@ukaachen.de; sipreuss@ukaachen.de; cneuschaefer@ukaachen.de; peterbirkholz@gmx.de).

C. Hey is with the Department of Phoniatrics and Pediatric Audiology, University of Frankfurt/Main, Frankfurt, Germany (email: Christiane.Hey@kgu.de).

R. Sader is with the Department of Oral, Cranio-, Maxillofacial and Facial Plastic Surgery, University of Frankfurt/Main, Frankfurt, Germany (email: r.sader@em.uni-frankfurt.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2312456

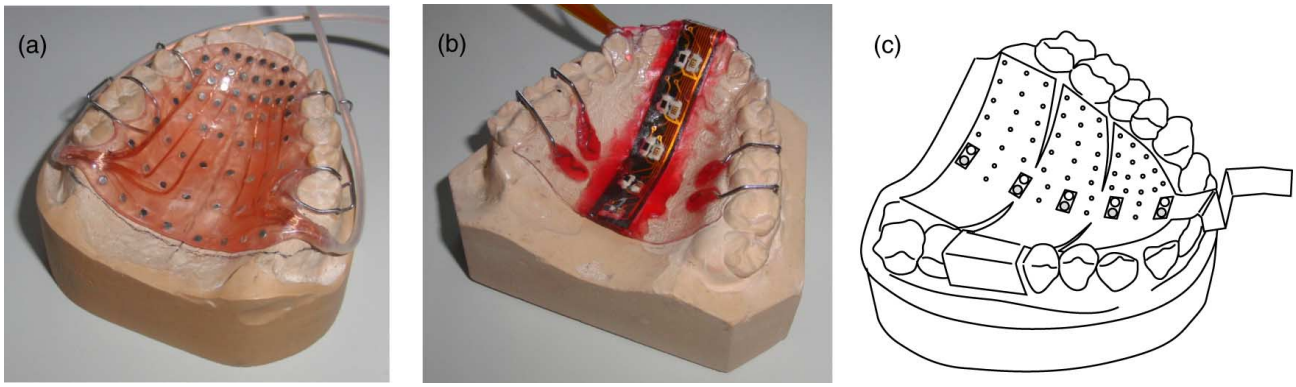

Fig. 1. (a) Typical Reading-type EPG palate with 62 contact sensors. (b) OPG palate with five optical distance sensors along the midline, mounted on a flexible circuit board. (c) Possible combination of OPG and EPG with both distance and contact sensors (contact sensors shown on one palate side only).

tour of the tongue from these contact or distance patterns, and whether the prediction would benefit from combining EPG and OPG data. A method for combining EPG and OPG measurements is currently under development in our group [16], [17] and a sketch of the new artificial palate is shown in Fig. 1(c).

To train and test the predictive models, we analyzed 3D MRI corpora of the vocal tract of sustained phonemes of two speakers, from which we extracted the tongue contours and the EPG and OPG patterns corresponding to the articulations. The model performance was assessed for the case that the models were trained and tested with the same speaker, and for the case that they were trained with one speaker and tested with the other.

II. Method

A. MRI Data

We used MRI corpora of two speakers. One corpus contained 3D scans of sustained phonemes of a male German speaker [18], of which we used the vowels /aː, eː, iː, oː, uː, ɛː, øː, yː, ɛ, ʊ, œ, a, I, Y, ɔ, ə, ɐ/ and the consonants /f, s, ʃ, ç, x, m, n, l/ (25 samples in total). Each phoneme was recorded with 18 sagittal slices of 3.5 mm thickness, 512×512 pixels per slice, and a pixel size of 0.59×0.59 mm^2. The acquisition took 21 s per phoneme. The other corpus contained 3D scans of sustained articulations of a male English speaker [19], of which we analyzed the vowels /I, ɛ, æ, ɒ, ʌ, ʊ, i, u, ɜ, ɑ, ɔ, ə, ɛː/ and the consonants /f, θ, s, ʃ, m, n, ɹ, l, x, ɬ, p, t/ (25 samples in total). Each 3D scan consisted of 26 sagittal slices of 4 mm thickness, 256×256 pixels per slice, and a pixel size of 1.1×1.1 mm^2. The average acquisition time per phoneme was 20 s.

B. Measurements

For each sample (phoneme) in the two corpora, we determined (1) the midsagittal contour of the tongue, (2) the linguopalatal contact pattern that would have been measured by EPG, and (3) five distances between the hard palate and the tongue that would have been measured by OPG.

In a first step, we traced the contours of the vocal tract in the midsagittal slice of each sample as shown in Fig. 2(b). The examination of these contours indicated that the head orientation was different for the two speakers during the scans, i.e, the head was rotated further backwards for the German speaker than for the English speaker. Based on the contours of the hard and (maximally raised) soft palate, we determined an angle by which all German samples were rotated to match the head orientation in the English samples before further processing.

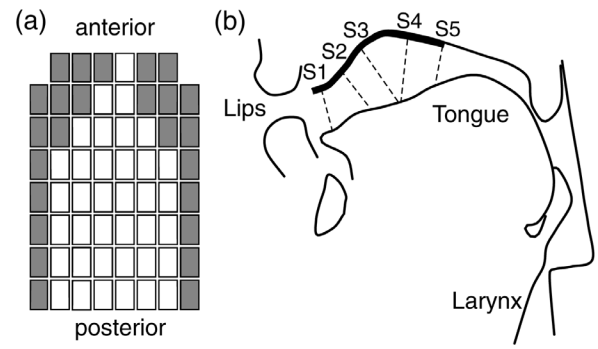

Fig. 2. (a) Typical EPG pattern of the fricative /s/. Electrodes with tongue contact are shown as grey boxes, and electrodes without contact as white boxes. (b) The OPG palate measures the distance from the sensors S1-S5 on the palate to the tongue along the optical axes of the sensors (dashed lines).

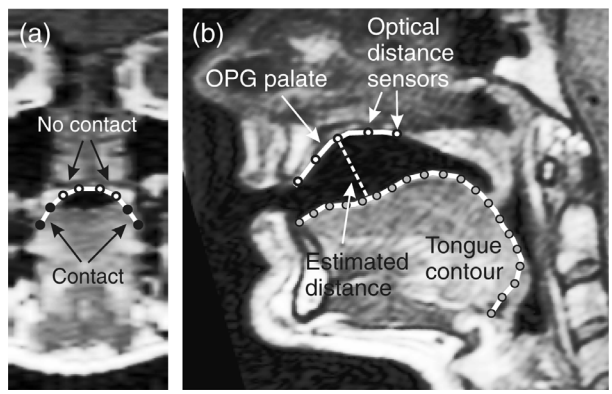

Fig. 3. (a) A coronal slice in the region of the hard palate, from which the state of one row of contact sensors of the virtual EPG palate was obtained. (b) Measurement of the palatolingual distances based on a virtual OPG palate. The tongue contour was represented by 20 equally spaced points.

For each sample, the contour of the tongue was represented by 20 points that were equally distributed between the tongue tip and the hyoid in the midsagittal slice, as shown in Fig. 3(b). The coordinates of the points were measured with respect to the most posterior point of the hard palate. To obtain the OPG data, five "virtual" distance sensors were equally distributed along the hard palate contour. The optical axes of the sensors

were assumed perpendicular to the palate contour, and the linguopalatal distances were measured along these axes as illustrated in Fig. 3(b).

EPG data were obtained by analyzing the coronal MRI slices at the positions of the eight electrode rows of a virtual Reading-type EPG palate. The positions of the electrode rows were defined according to the scheme in [11], i.e., the first row was located at the anterior end of the EPG palate, the last row was located at the posterior end of the EPG palate, and the remaining rows were arranged so that the spacing between the front four rows was about half that of the back four rows. The electrodes in each row were positioned equally spaced along the palatal vault from the left to the right gingival margin. Fig. 3a shows one of the electrode rows in a coronal slice of the vowel /ɛ:/. Contacted electrodes were determined by visual inspection, yielding for each sample a binary contact pattern as in Fig. 2a. Prior to their use with the predictive models, we reduced the dimensionality of the contact patterns. This was done for two reasons: (1) The binary contact data in EPG patterns are highly correlated due to the limited degrees of freedom of the tongue, and (2) the number of available training samples is rather small. Nguyen [20] showed that four low-frequency coefficients of the discrete 2D cosine transform of the EPG patterns represent most of the essential information they contain and capture several characteristics that are relevant from an articulatory point of view. Therefore, these four coefficients were used here as compact representation of the patterns for the prediction models. According to [20], the four coefficients indicate (1) the scaled sum of the activated electrodes in an EPG pattern, (2) the left-right asymmetry in a pattern in terms of the difference in the number of activated electrodes between the left and right sides of the palate, (3) the arrangement of contacts along the front-back dimension, and (4) whether there are more activated electrodes along the lateral sides or along the median line.

C. Predictive Model

To predict the coordinates of the 20 points representing the tongue contour in the midsagittal plane, we propose to use multiple linear regression. Hence, given a vector $(c_1\ c_2 \ldots c_K)$ of sensor data (OPG or EPG data), we assume that each point (x_j, y_j) of the tongue contour ($j = 1 \ldots 20$), can be expressed as

$$x_j = a_{j,0} + \sum_{k=1}^{K} a_{j,k} \cdot c_k \quad \text{and} \quad y_j = b_{j,0} + \sum_{k=1}^{K} b_{j,k} \cdot c_k \quad (1)$$

with $a_{j,k}$ and $b_{j,k}$ being the parameters of the model. The parameters were estimated by fitting the model to the samples derived from MRI in the least-squares sense.

We examined three settings with respect to the vector of sensor data. In the first setting, the vector contained the $K = 4$ EPG indices, i.e., the tongue contour was predicted from the contact patterns only. In the second setting, the vector contained only the $K = 5$ palatolingual distances as measured by the optical sensors. In the last setting, the vectors of EPG and OPG data were concatenated, yielding a vector with $K = 9$ elements.

In settings two and three, the measurement d_1 of the most anterior distance sensor was weighted with the factor $1/(1 + d_1)$, with d_1 in cm, during training and testing. This was done because in some samples the tongue tip was behind the optical axis of this distance sensor so that the sensor measured the distance to the mouth floor instead of to the tongue tip. If we assume two samples with nearly the same tongue shape, but with a more anterior tongue tip in one sample than in the other, d_1 could differ considerably between the samples, because the anterior sensor measures the distance to the tongue tip in one sample, and to the floor of the mouth in the other sample. This is a non-linear behaviour that may degrade the performance of linear predictive models. To reduce this problem, the weighting of d_1 was introduced to penalize high d_1 values, i.e, when the tongue tip is behind the sensor axis, while low d_1 values are still trusted.

D. Evaluation

The performance of the predictive models was assessed under four conditions: (1) With leave-one-out cross-validation using only the German samples, (2) with leave-one-out cross-validation using only the English samples, (3) training on all German samples and testing on all English samples, and (4) training on all English samples and testing on all German samples. For the latter two inter-speaker evaluations, we implemented a method to adapt the predicted contours to the respective other speaker by compensating the differences in the shapes of the hard palates. Assume that (x_0, y_0) is a predicted tongue point, and that y_{train} and y_{test} are the vertical positions of the palate contours of the training speaker and the test speaker at the horizontal position x_0. In this case, y_0 was adjusted to

$$y_0' = y_0 + (y_{\text{test}} - y_{\text{train}}) \cdot e^{-\|y_0 - y_{\text{train}}\|/T}. \quad (2)$$

Hence, when the tongue was close to or touched the palate ($\|y_0 - y_{\text{train}}\| \to 0$), the difference in palate shape was fully compensated to correctly represent vocal tract closures or critical constrictions. For lower tongue positions, the compensation was exponentially reduced for smoother tongue shapes. The value of the decay constant T is not critical and was set to 4 cm.

To test a model with a given sample, the virtual sensor data of the sample were used to predict a tongue contour that was then compared with the measured tongue contour for the sample. Therefore, for each of the 20 points defining the measured contour, the closest Euclidian distance d_i ($i = 1 \ldots 20$) to the predicted contour was calculated. The error was defined as the mean value of these distances, i.e., $(\sum_{i=1}^{20} d_i)/20$.

III. Results and Discussion

The results of this study are summarized in Figs. 4 and 5. Fig. 4 shows that the prediction errors were generally lower for the intra-speaker evaluations than for the inter-speaker evaluations. For each training-test condition, two-tailed two-sample paired Student's t-tests to the 5% significance level were performed to test whether the errors differed between the pairs of settings. Bonferroni correction was applied to account for the three tests per condition.

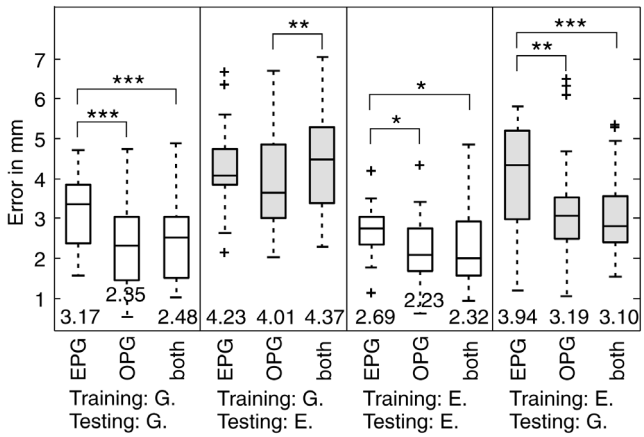

Fig. 4. Error distributions of the predicted tongue contours for the different combinations of training and test samples (G = German speaker; E = English speaker), separated by the setting (EPG, OPG, and combined EPG and OPG). Each boxplot represents 25 samples. Mean values are written below the box plots. Significant differences are indicated as follows: * $p < 0.05$; * $p < 0.01$; *** $p < 0.001$.

For the intra-speaker conditions, the prediction errors based on EPG data were significantly higher than the errors based on OPG or combined EPG and OPG data. The reason is probably that EPG provides no information about the tongue at all when it does not touch the hard palate. In this case, the predicted tongue shape is always the same albeit the real tongue shape may differ substantially. In each of our corpora, there were actually five samples without any EPG contact. OPG, on the other hand, can discriminate tongue shapes even when there is no palatolingual contact. The prediction errors based on combined EPG and OPG data were not significantly different from that based on OPG data only. Hence, either the additional information in the EPG patterns is redundant or the number of training samples was too low to allow the models to acquire sufficient generalization capabilities for the number of nine sensor variables in the third setting. Hence, future studies should assess the model performance with a higher number of MRI samples. For example, 64 training samples were required in the related study [5] to achieve a minimum of the prediction error.

Under the two inter-speaker conditions, there was no consistent relation between the prediction errors and the three settings. The generally higher errors for these conditions indicate that a more sophisticated method for speaker adaptation is necessary. For future work, a linear transformation for speaker adaptation as proposed in [7] might to be a good starting point.

Examples of measured and predicted contours in both corpora based on OPG data are shown in Fig. 5. For each of the two shown conditions (intra-speaker vs. inter-speaker), the test samples with the lowest error (best examples), the median error (typical examples), and the highest error (worst examples) are presented. Apart from the worst case examples, we note that the tongue contour in the oral cavity, where the actual OPG distances were measured, was reproduced quite well, even under the inter-speaker condition, but was somewhat worse in the pharyngeal region. The supplemental material contains a figure that shows the prediction error individually for the 20 points that define the tongue contour. The error is clearly lower for points 1-10

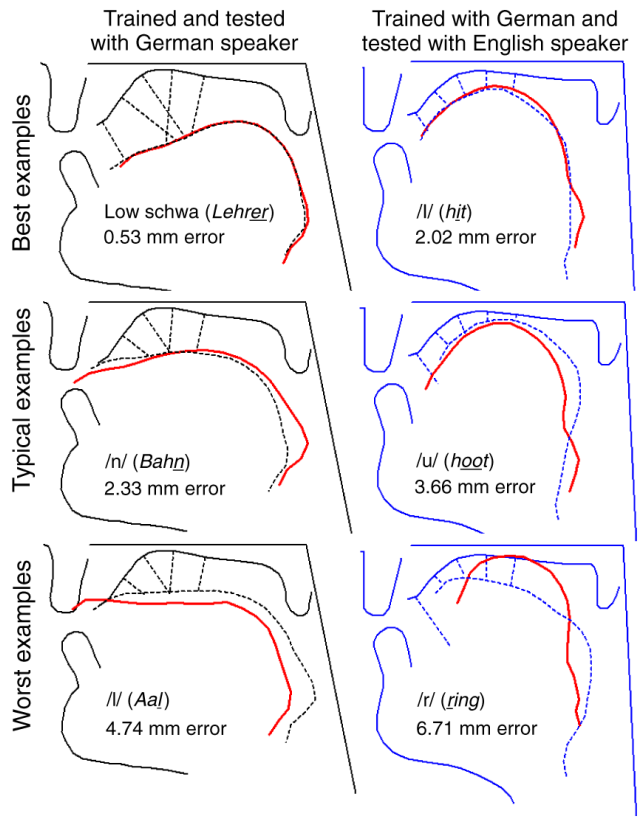

Fig. 5. Best, typical (median), and worst tongue contour predictions based on OPG data for an intra-speaker and an inter-speaker condition. The predicted contours are drawn as solid lines, and the measured contours as dashed lines. The straight dashed lines indicate the optical axes of the distance sensors.

(oral region) than for points 11-20 (pharyngeal region). This indicates that the shape of the posterior part of the tongue is not completely predictable from the measured anterior part of the tongue, but has more degrees of freedom. The English phoneme /r/ was predicted with the highest error of all (when the model was trained with the German speaker). The cause for this is not only the need for better speaker adaptation and a larger sample size for training, but also that the most anterior sensor measured the distance to the floor of the mouth instead of to the tongue tip, as discussed in Section II-C. In general, it remains an open question whether the inter-speaker predictions are less successful because of the different languages of the corpora or because of the different anatomy of the speakers.

Despite these limitations, the predicted contours appear sufficiently realistic for animating tongue movements for visual feedback in speech therapy, for example, especially when we consider that the important anterior part of the tongue is predicted more accurately than the posterior part. In contrast to previous approaches to tongue contour prediction based on EMA, the proposed methods EPG and OPG are more convenient for repeated use by subjects or patients. Furthermore, MRI-derived tongue contours are more detailed than previously used US-derived contours, but larger MRI corpora are needed for more training samples. The tongue shapes and the simulated EPG/OPG patterns are available from www.vocaltractlab.de (link to supplemental materials).

REFERENCES

[1] O. Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher," *Comput. Assisted Lang. Learn.*, vol. 25, no. 1, pp. 37–64, 2011.

[2] K. Richmond and S. Renals, "Ultrax: An animated midsagittal vocal tract display for speech therapy," in *Interspeech 2012*, Portland, USA, 2012, pp. 74–77.

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

[4] 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 Commun. Sci. Disorders*, vol. 30, pp. 2–25, 2003.

[5] T. Kaburagi and M. Honda, "Determination of sagittal tongue shape from the positions of points on the tongue surface," *J. Acoust. Soc. Amer.*, vol. 96, no. 3, 1994.

[6] C. Qin, M. A. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," in *Interspeech 2008*, Brisbane, Australia, 2008, pp. 2306–2309.

[7] C. Qin and M. A. Carreira-Perpiñán, "Adaptation of a predictive model of tongue shapes," in *Interspeech 2009*, Brighton, U.K., 2009, pp. 772–775.

[8] C. Qin and M. Carreira-Perpiñán, "Reconstructing the full tongue contour from EMA/X-ray microbeam," in *IEEE International Conf. Acoustics Speech and Signal Processing (ICASSP 2010)*, 2010, pp. 4190–4193.

[9] P. Badin, E. Baricchi, and A. Vilain, "Determining tongue articulation: From discrete fleshpoints to continuous shadow," in *Eurospeech 1997*, Rhodes, Greece, 1997, pp. 47–50.

[10] S. Preuß, C. Neuschaefer-Rube, and P. Birkholz, "Real-time control of a 2D animation model of the vocal tract using optopalatography," in *Interspeech 2013*, Lyon, France, 2013, pp. 997–1001.

[11] W. Hardcastle, W. Jones, C. Knight, A. Trudgeon, and G. Calder, "New developments in electropalatography: A state-of-the-art report," *Clinical Linguist. Phonet.*, vol. 3, no. 1, pp. 1–38, 1989.

[12] C.-K. Chuang and W. S. Wang, "Use of optical distance sensing to track tongue motion," *J. Speech Hearing Res.*, 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 Linguist. Phonet.*, vol. 3, no. 4, pp. 359–375, 1989.

[14] A. A. Wrench, A. D. McIntosh, C. Watson, and W. J. Hardcastle, "Optopalatograph: Real-time feedback of tongue movement in 3D," in *5th Int. Conf. Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, 1998, pp. 305–308.

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

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

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

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

[19] I. Steiner, K. Richmond, I. Marshall, and C. D. Gray, "The magnetic resonance imaging subset of the mngu0 articulatory corpus," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. EL106–EL111, 2012.

[20] N. Nguyen, "EPG bidimensional data reduction," *Int. J. Lang. Commun. Disorders*, vol. 30, no. 2, pp. 175–182, 1995.