Optical Sensor Calibration for Electro-Optical Stomatography

Simon Preuß, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden

simon_preuss@tu-dresden.de, peter.birkholz@tu-dresden.de

Abstract

We are currently developing a technology called "electro-optical stomatography" to measure and visualize articulatory movements within the vocal tract using electrical contact sensors and optical proximity sensors. To measure tongue movements with the optical sensors in this system, a mapping between the raw sensor values and actual tongue positions has to be determined. This mapping is non-linear and different for every tongue and sensor. The lack of an accurate, reliable calibration method has so far prevented wide-spread use of optical measurements within the vocal tract. Here, we present a calibration method based on a multi-linear regression model that maps the sensor value at a single distance of $0\,\mathrm{mm}$ to calibration values at 0, 5, 10, 15, 20, 25, and $30\,\mathrm{mm}$. The coefficients of the model are determined by a least-squares regression in 25 training data sets (recorded with 5 subjects and 5 sensors). Evaluation in a leave-one-out cross-validation and on five more data sets recorded with another, different subject on 5 additional sensors yields very good results with maximum median position errors close to $1\,\mathrm{mm}$. The calibration of the optical sensors can therefore be semi-automatically accomplished based on a single, easily obtainable measurement during direct tongue contact.

Index Terms: optical distance sensing, articulography, electro-optical stomatography, speech movements, articulation

1. Introduction

Speech cannot be analyzed without instrumental technology to record and visualize speech data. Most researchers in the field equate speech data with acoustic data recorded by microphones. However, speech is produced by the coordinated movements of the articulators: mainly the lips, velum, jaw, and, above all, the tongue. These articulatory movements can be just as informative about the speech as the audio signal itself, or even more so. Measuring articulatory data is more difficult than recording sound though, because the articulators are mostly hidden from view inside the mouth cavity and the throat (i. e., the vocal tract). A number of technologies have been employed in the past to acquire these data (see [1] for a review), among them sonography, electro-magnetic articulography (EMA), X-ray microbeam, and electromyography (EMG). Though a lot of progress was made using these techniques, they still remain impractical for many applications due to their cost, complexity, necessity of specialized personnel, or strenuous application for the subject. The current lack of a system that is easy to use, sufficiently precise, portable, and versatile enough to measure the entire range of articulation motivated us to develop a new technique called "electro-optical stomatography" (EOS) that can help to fill this technological gap (see, e. g., [2, 3, 4, 5, 6, 7, 8]).

2. Electro-optical stomatography

Electro-optical stomatography (EOS) is basically a combination of two established techniques: electropalatography (EPG) and optopalatography (OPG), also known as glossometry. EPG was first introduced in the 1970s and since has become increasingly popularized within the speech research community as well as in clinical applications [9]. With EPG, the time-varying contact pattern of the tongue on the hard palate is measured by means of an array of electrical contact sensors that are arranged on a plastic plate fitted to a subject's individual palate shape (a "pseudopalate"). These sensors register a small voltage that is applied to the subject, when the tongue touches the sensors, thus outputting a binary value that corresponds to "contact" or "no contact", respectively. Different EPG systems were and are in existence that primarily differ in number and arrangement of the contact sensors. The two currently commercially available systems use 62 (see [10]) and up to 126 (see [11]) sensors, respectively. Our system uses 124 contact sensors that are fixed to the palate using a flexible circuit board. This simplifies the manufacturing process significantly and thus reduces the total cost of the system. The contact sensors are also switched through in sequence on the pseudopalate by four small analog multiplexers embedded between the sensor board and the plastic base plate. This reduces the number of wires that need to exit the mouth, which improves the wearing comfort of the pseudopalate.

As EPG is intrinsically "blind" to all tongue movements that do not contain palate contact (such as the movements necessary to produce open or back vowels), we looked for a complementary technology and found optopalatography (OPG) to be a suitable addition: OPG was first presented by Chuang and Wang in 1978 [12] and further developed by Fletcher et al. in the 1980s (see, e. g., [13, 14]) and by Wrench et al. in the 1990s (see, e. g., [15, 16, 17]). This technology also uses a pseudopalate, but in this case it carries optical sensors that measure the distance between the tongue and the hard palate: Infrared light is emitted by the sensors that is then diffusely reflected by the tongue surface. The sensors measure the intensity of the reflected light, which is related to the distance of the reflecting surface from the receiver. In our system, we use five miniature laser diodes that emit narrow infrared laser beams as sources and five phototransistors with matching wavelength sensitivity as receivers. These sensors are arranged along the sagittal midline of the pseudopalate. A sixth sensor, consisting of one laser diode and two receiving phototransistors, is placed in front of the upper incisors and measures the lip opening and protrusion (see [2] and [6] for details). Figure 1 shows an assembled EOS pseudopalate. By combining the palato-lingual contact information of the EPG measurements with the additional knowledge on the sagittal tongue contour gleaned from the OPG measurements, the entire articulatory space (with respect to the tongue movements) is covered by an EOS system.

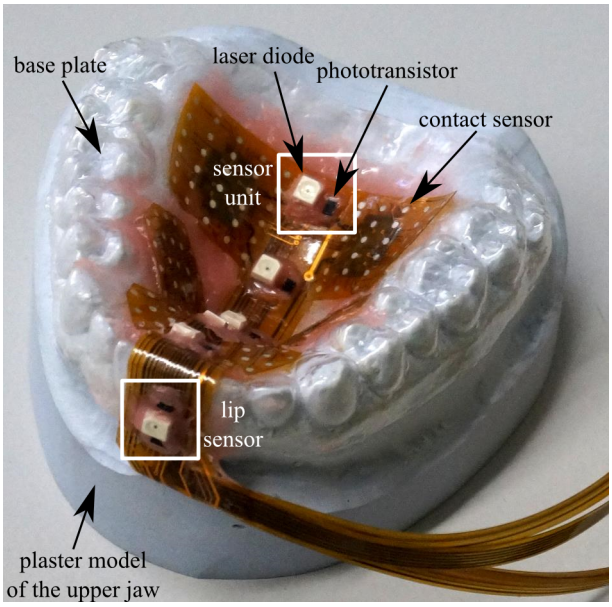

Figure 1: An assembled pseudopalate for EOS measurements.

The rate at which an entire data frame of contact and distance data is gathered is 100 Hz and therefore sufficient for real-time analysis of continuous speech. The control unit connected to the pseudopalate measures only about 16 cm x 10 cm, is powered by a 9 V battery and thus highly portable. The control unit sends the measurement data to a PC or laptop using a standard serial connection, so no sophisticated equipment is needed for data acquisition. A custom Windows software is used to record and display the time series of measured data offline and in a real-time view and in sync with the audio signal.

3. Calibration of the optical sensors

In our system, the distance of the reflector from an optical sensor is mapped to a 12-bit analog-to-digital-converted (ADC) sensor value between 0 and 4095 (see Figure 2). Because of the complex optical relationships, formulating this mapping in a single analytic expression is impossible (or at least impractical). Chuang and Wang tried in [12] and came very close, but ultimately their simplifying assumption that the tongue surface behaves like an ideal Lambert reflector could not be successfully proven in our experiments and the function described in their paper could not be fitted sufficiently well to our observed data. This is because real-life tongues obviously differ slightly from one another and show different reflective properties that need to be compensated by the sensor calibration on a case-by-case basis. Also, the sensors themselves have slightly different characteristics because of their manual assembly and embedding. Figure 2 (a) illustrates the differences of calibration characteristics measured with different tongues and different sensors (see below) and demonstrates the importance of an individually adjusted calibration. Our first approach was to calibrate the sensors by acquiring 7 sensor values at 7 different well-defined distances (0, 5, 10, 15, 20, 25, and 30 mm) in vitro (see [3]) and linearly interpolating between these points (distances larger than 30 mm are generally irrelevant for continuous articulation). While this yields a good approximation of the actual mapping, it is not possible to repeat the calibration process once the pseudopalate has been assembled, because well-defined distances cannot be ensured in situ.

However, we postulate that the different reflective properties of different tongues and the slight variations across different sensors are sufficiently represented in the sensor value x_0 at a distance of 0 mm, i. e., during direct tongue contact, and that there exists a (non-linear) relationship between x_0 and the sensor values x_i at the other six distances d_i (d_1 - d_6: 5, 10, 15, 20, 25, and 30 mm). If this assumption holds and the mapping from x_0 to x_i can be found, all seven calibration values can be acquired with a single measurement that can even be performed after the assembly of the pseudopalate (i. e., it can be repeated to adapt to intra-individual time-varying differences of the reflective properties of the tongue).

The function f_i that relates x_0 to x_i could possibly be very complex, too. However, we can expand f_i into a Taylor series and truncate it to the power of two, i. e., the second-order Taylor polynomial as shown in Eq. (1). This is a non-linear approximation of the unknown function f_i at the point p_i. If we expand these terms and sort by the power of x we obtain (2). As p_i is a specific value, we can further simplify the expression to (3).

$$f_i(x) = f_i(p_i) + \frac{f_i'(p_i)}{1!}(x - p_i) + \frac{f_i''(p_i)}{2!}(x - p_i)^2 \quad (1)$$

$$= \underbrace{f_i(p_i) - \frac{f_i'(p_i)}{1!}p_i - \frac{f_i''(p_i)}{2!}p_i^2}_{a_{i,0}} + \underbrace{\left(\frac{f_i'(p_i)}{1!} - \frac{f_i''(p_i)}{2!}2p_i\right)}_{a_{i,1}} x + \underbrace{\frac{f_i''(p_i)}{2!}}_{a_{i,2}} x^2 \quad (2)$$

$$= a_{i,0} + a_{i,1}x + a_{i,2}x^2 \quad (3)$$

We now would have to determine the scalar coefficients $a_{i,0}$, $a_{i,1}$, and $a_{i,2}$ so that $f_i(x_0)$ becomes x_i. Because a single exact solution of this problem that holds for all tongues and sensors with just a single set of coefficients for each distance is not possible, we need to find optimal sets that yield a good approximation $f_i(x_0) = \hat{x}_i \approx x_i$. To that end, we need a number of known tuples (x_0, x_i) for each distance d_i to set up an overdetermined set of (in the coefficients) linear equations as in (4), where $\mathbf{x_i}$ is a column vector containing sensor values measured at distance d_i, $\mathbf{X_0} = (\mathbf{1},\ \mathbf{x_0},\ \mathbf{x_0}^2)$ (where $\mathbf{x_0}$ is a column vector of sensor values at a distance of 0 mm and $\mathbf{x_0}^2$ denotes the element-wise square of $\mathbf{x_0}$), and $\mathbf{a_i} = (a_{i,0},\ a_{i,1},\ a_{i,2})^T$.

$$\mathbf{x_i} = \mathbf{X_0} \cdot \mathbf{a_i} \quad (4)$$

This system of equations is set up for every distance d_i and solved for $\mathbf{a_i}$ in a least-squares optimal sense using a standard QR decomposition algorithm in MATLAB. This eventually yields one set of coefficients for each distance at which a sensor value needs to be calculated.

3.1. The training data

In order to determine the coefficients, training data are needed to set up the overdetermined system of equations. The training data are digital sensor values measured at seven distances $d_i = \{0, 5, 10, 15, 20, 25, 30\}$ (all in mm) with five male subjects (age 29 to 62) and five different sensors. In order to ensure well-defined and stable distances, we used spacers that have a minimal impact on the sensor value compared to a measurement at the same distance without spacers. In [8] we presented and evaluated a number of different configurations of such spacers

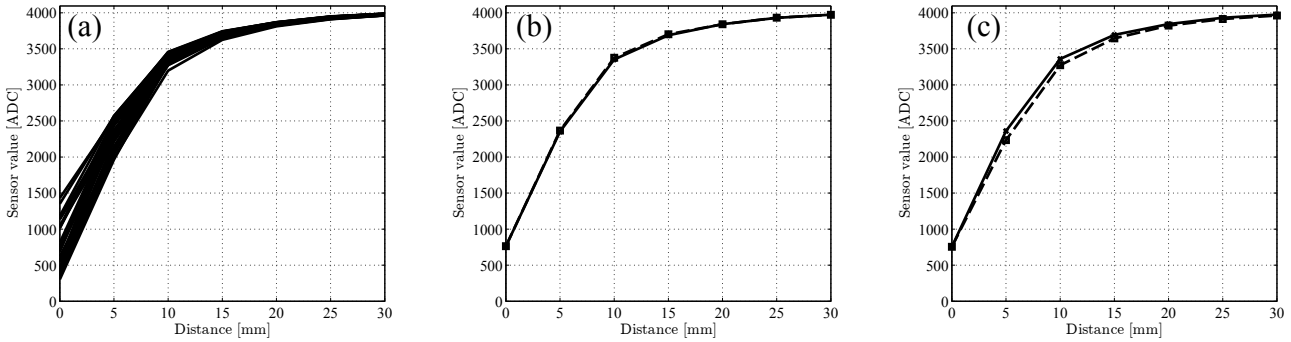

Figure 2: (a) Training data consisting of 25 calibration characteristics measured with 5 tongues and 5 sensors. (b) Best overall fit (RMSE: $0.1\,\mathrm{mm}$) and (c) worst overall fit (RMSE: $1.1\,\mathrm{mm}$) of calculated calibration points (solid line, cross markers) versus measured data (dashed line, square markers) in the leave-on-out cross-validation paradigm.

of which we employed the best one in this study. The spacers are plexiglass tube segments, inner diameter $34\,\mathrm{mm}$, of the desired lengths, whose inner walls are lined with black cardboard to avoid stray reflections. They are open on one end, so they can be placed over the sensors, and on the other end capped with a plexiglass grid (width of the bracers: $0.5\,\mathrm{mm}$; width of the gaps: $4.5\,\mathrm{mm}$). This grid keeps the tongue from bending into the tube, which would reduce the distance between the sensor and the tongue surface. These spacers were placed between the subject's tongue and the sensor board, which in turn was fixed to a stable, black surface to avoid displacement of the sensor and stray reflections. The sensor value associated with the respective distance (i. e., length of the spacer) was then determined as the mean value of three repetitions of a measurement averaged over $500\,\mathrm{ms}$. This process was repeated with each subject at all five sensors. The training corpus therefore consists of a total of 25 trials of measurements at 7 distances each.

3.2. Evaluation

By excluding a single trial n from the training data before solving for $\mathbf{a_i}$, this trial can then be used to test the coefficients found on the rest of the data set by calculating $\hat{x}_i = (1,\ x_{n,0},\ x_{n,0}^2,) \cdot \mathbf{a_i}$ and determining the difference between the measured and calculated points $\hat{x}_i - x_{n,i}$ and converting it to mm. The entire process is done 25 times, each time excluding another trial n (a scheme commonly known as leave-one-out cross-validation). The results of all 25 rounds are summarized in Figure 3. The median error is less than $0.1\,\mathrm{mm}$ at all distances while the maximum error is $1.8\,\mathrm{mm}$ at an actual distance of $30\,\mathrm{mm}$. Figure 2 (b) and (c) show two representative examples of generated points versus measured data from the leave-one-out cross-validation.

A *second* set of evaluation data was recorded with the same setup as the training data except that we used an additional 6th subject and 5 different sensors. The calibration points where calculated using a set of coefficients determined on the first set of 25 trials and the respective x_0 of the new set. The resulting errors are presented in Figure 4. The median error was less than $1\,\mathrm{mm}$ at most distances except at $10\,\mathrm{mm}$ where it amounted to $1.17\,\mathrm{mm}$, the maximum error was $1.95\,\mathrm{mm}$ at an actual distance of $25\,\mathrm{mm}$.

3.3. Discussion

The results in the leave-one-out cross-validation setting are significantly better and more consistent than the results on the second evaluation set. This is due to the fact that by leaving only one trial out, data from the tongue that was used in this trial is still present in the training set, albeit recorded with another sensor. Analogously, the one sensor used in the left-out trial is still represented in the training set by measurements with different tongues. Therefore, the results from the second evaluation set with the entirely new speaker are to be considered indicative for the true quality of generalization and representative for the error we expect to find in real-world applications. By adding trials with more sensors and more subjects to the training corpus, we would expect to further improve the results.

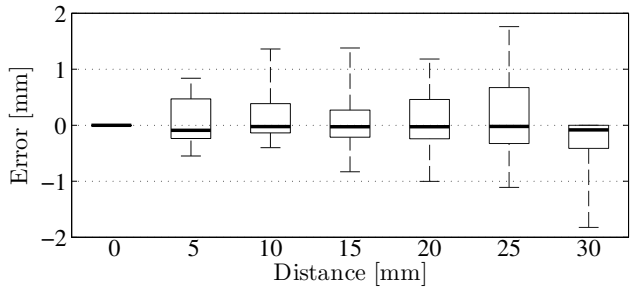

Figure 3: Error in mm at each distance across the 25 samples. The thick line is the median, the edges of the boxes mark the 25th and 75th percentiles, and the whiskers stretch to the most extreme points.

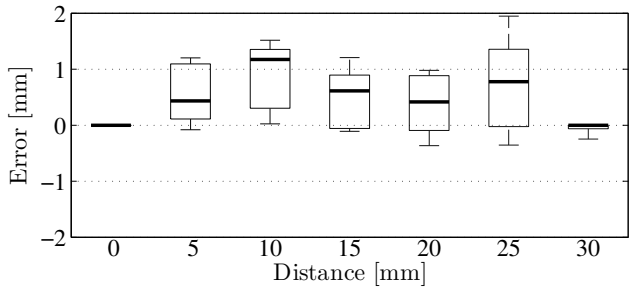

Figure 4: Error in mm at each distance across the 5 trials recorded with a subject and sensors not part of the training set.

The results also show that the error can become generally larger at greater distances than at closer distances. Because of the decreasing $\frac{ADC}{\mathrm{mm}}$ resolution between the calibration points with increasing distance (see Figure 2), a small error in the calculated sensor value becomes an increasingly larger error in mm at greater distances. If the resolution could somehow be increased at these distances, maybe even at the expense of the resolution at closer distances, this effect might be compensated. The measured calibration points during tongue contact x_0 in the training data set also varied significantly across subjects and sensors between about 300_{ADC} to 1400_{ADC}. While we attribute small differences in the sensor values to the varying optical properties between subjects and sensors, this range seems too large to be explained solely by these variables. In a future study we would therefore ensure that the subjects apply consistent pressure to the sensor when their tongues are placed directly on it. This way we could evaluate if Chuang and Wang's observation, that the sensors are capable to measure "negative distances" (see [12]), could indeed be used to register the force with which the tongue is pressed against the palate. A less scattered distribution of x_0 values might also further reduce the error when calculating the calibration points. But since we were looking for a calibration scheme that would allow a simple "update" of the calibration even after the pseudopalate is assembled and well-defined conditions can no longer be ensured, we chose to use the entire, scattered distribution of x_0 values as we cannot control how forceful a subject presses its tongue against the sensor during the intended application scenario.

4. Examples of measured contours

In order to test the plausibility of tongue contours obtained with a regression-based calibration, we recorded synchronized audio and EOS data of 10 realizations each of five sustained vowels (/aː, eː, iː, oː, uː/). Data from this speaker were part of the training corpus of the calibration model but the sensors mounted on the pseudopalate were different from both the ones used in the training and the evaluation corpus. Using the phonetics software PRAAT [18], we segmented the recordings and extracted the middle sections of each realization, keeping the section length approximately constant at about $300\,\mathrm{ms}$. In the EOS software, we then averaged the EOS data over these intervals. The resulting mean, lowest and highest contours of the five vowels are presented in Figure 5. The tongue shapes are generally plausible and in line with the phonetic features height and backness. There is also only a very small difference between the lowest and highest shapes within the 10 repetitions of each vowel which indiciates a high reproducibility within a series of measurements.

5. Conclusions

As the results on the evaluation set and the contours given in Figure 5 show, the automatic calibration using Equation (3) and the coefficients determined in this study perform sufficiently well for realistic and meaningful measurements. The small errors are outweighed by the ability to quickly adapt to varying optical conditions during a measurement and the fact that a subject no longer has to do a calibration trial at several distances with each sensor before the pseudopalate is assembled. In future works we will focus on further reducing the calibration error by increasing the size of the training data and reducing the variance of the x_0 distribution by controlling the level of force the tongue exerts on the sensors. Another important future experiment is to perform sonography in conjunction with EOS measurements to evaluate the absolute precision of our system in situ. Furthermore, the measurement technique so far relies on a single sensor reading. In future studies, we intend to examine if sensor readings of neighboring sensors can be used to improve the precision or stability versus angular displacement of the tongue surface with respect to the optical sensor axes.

6. Acknowledgments

This work was funded by the German Research Foundation (DFG), grants BI 1639/1-1 and BI 1639/1-2.

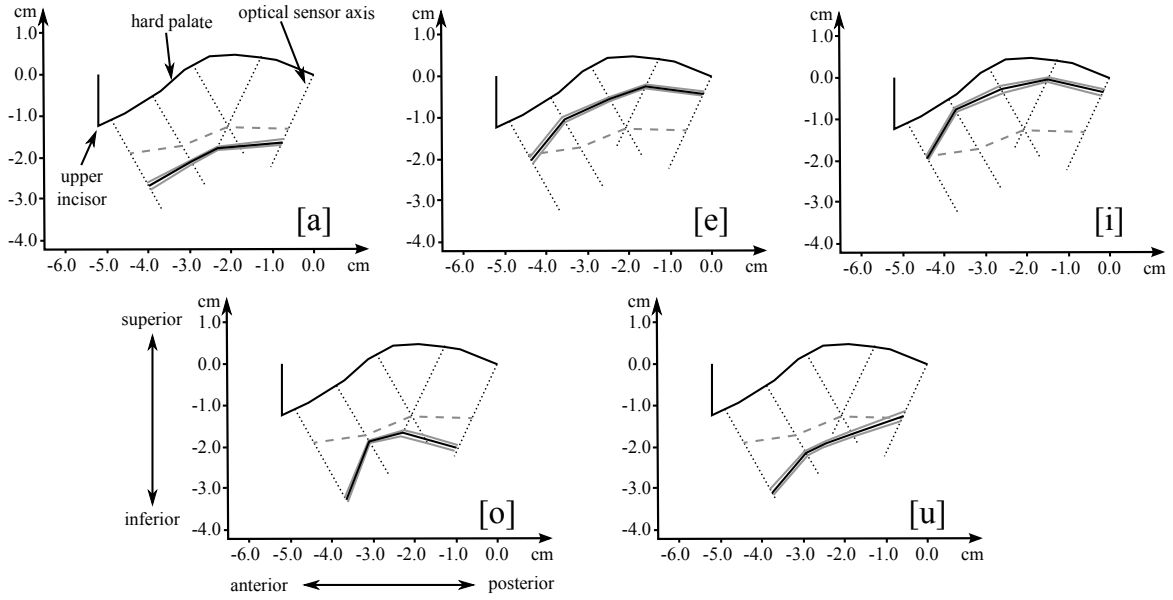

Figure 5: Tongue shapes of five vowels obtained with a regression-based calibration. The solid black line is the mean shape (of 10 repetitions each), gray solid lines are highest and lowest shapes, and dashed gray line is the contour of the neutral vowel /ə/ as a reference.

7. References

[1] 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.

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

[3] ——, "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.

[4] P. Birkholz, P. Dächert, and C. Neuschaefer-Rube, "Advances in combined electro-optical palatography," in *Proc. of Interspeech 2012*, Portland, Oregon, USA, 2012, pp. 703–706.

[5] 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 *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2013*, M. Wolff, Ed. Dresden, Germany: TUDPress, 2013, pp. 144–151.

[6] ——, "Real-time control of a 2D animation model of the vocal tract using optopalatography," in *Proc. of Interspeech 2013*, Lyon, France, 2013, pp. 997–1001.

[7] R. Mumtaz, S. Preuß, C. Neuschaefer-Rube, C. Hey, R. Sader, and P. Birkholz, "Tongue contour reconstruction from optical and electrical palatography," *Signal Processing Letters, IEEE*, vol. 21, no. 6, pp. 658–662, June 2014.

[8] S. Preuß and P. Birkholz, "Fortschritte in der Elektro-Optischen Stomatographie," in *Accepted for: Elektronische Sprachsignalverarbeitung 2015*. Dresden, Germany: TUDPress, 2015.

[9] A. A. Wrench, "Advances in EPG palate design," *International Journal of Speech-Language Pathology*, vol. 9, no. 1, pp. 3–12, 2007.

[10] W. J. Hardcastle, F. E. Gibbon, and W. Jones, "Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders," *International Journal of Language & Communication Disorders*, vol. 26, no. 1, pp. 41–74, 1991.

[11] S. Fletcher, B. Sparks, and J. Tanner, "Palatometer and nasometer apparatus," Dec. 13 2005, US Patent 6974424.

[12] C.-K. Chuang and W. S.-Y. Wang, "Use of optical distance sensing to track tongue motion," *J Speech Hear Res*, vol. 21, no. 3, pp. 482–496, 1978.

[13] J. E. Flege, S. G. Fletcher, M. J. McCutcheon, and S. C. Smith, "The physiological specification of American English vowels," *Language and Speech*, vol. 29, no. 4, pp. 361–388, 1986.

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

[15] A. Wrench, A. D. McIntosh, and W. J. Hardcastle, "Optopalatograph (OPG): A new apparatus for speech production analysis," in *Proc. of the Fourth International Conference on Spoken Language (ICSLP), 1996.*, vol. 3. IEEE, 1996, pp. 1589–1592.

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

[17] A. A. Wrench, A. D. McIntosh, C. Watson, and W. J. Hardcastle, "Optopalatograph: Real-time feedback of tongue movement in 3D," in *5th International Conference on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, 1998, pp. 1867–1870.

[18] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.