INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Towards Minimally Invasive Velar State Detection in Normal and Silent Speech

Peter Birkholz[1], *Petko Bakardjiev*[1], *Steffen Kürbis*[1], *Rico Petrick*[2]

[1]Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
[2]Linguwerk GmbH, Dresden, Germany

peter.birkholz@tu-dresden.de

Abstract

We present a portable minimally invasive system to determine the state of the velum (raised or lowered) at a sampling rate of 40 Hz that works both during normal and silent speech. The system consists of a small capsule containing a miniature loudspeaker and a miniature microphone. The capsule is inserted into one nostril by about 10 mm. The loudspeaker emits chirps with a power band from 12-24 kHz into the nostril and the microphone records the signal reflected from the nasal cavity. The chirp response differs between raised and lowered velar positions, because the velar position determines the shape of the nasal cavity in the posterior part and hence its acoustic behaviour. Reference chirp responses for raised and lowered velar positions in combination with a spectral distance measure are used to infer the state of the velum. Here we discuss critical design aspects of the system and outline future improvements. Possible applications of the device include the detection of the velar state during silent speech recognition, medical assessment of velar mobility and speech production research.

Index Terms: velar state detection, silent speech

1. Introduction

The position of the velum determines whether or not the nasal cavity is acoustically coupled to the vocal tract. With a lowered velum, speech sounds are produced nasalized and with a raised velum they are produced non-nasalized. The possibility to detect and track the velar position has many potential applications. One application is silent speech recognition, i.e., the automatic recognition of speech that is produced silently without acoustic excitation of the vocal tract [1]. Existing techniques for silent speech recognition employ for example electromyography with surface electrodes on the face and neck [2, 3] or ultrasonography of the tongue in combination with video recordings of the lips [4, 5] as sensory data. However, neither of these modalities provides reliable information about the velar state, making a distinction of nasalized and non-nasalized speech sounds difficult. Hence, a method for minimally invasive velar state detection could complement these techniques for silent speech recognition. A medical application for velar state detection is the therapy of velopharyngeal dysfunction by biofeedback [6, 7]. Velopharyngeal dysfunction is characterized by nasal speech and is for example seen in patients with a cleft palate or certain neuromuscular problems. In basic research, velar state detection allows to analyze how velar movements are coordinated in conjunction with the other articulators to inform theories of speech motor control [8, 9].

Existing methods for velar state detection can be divided into general imaging techniques like MRI, X-ray, and nasal endoscopy on the one hand [10], and into dedicated methods on the other hand. The advantage of dedicated methods is that the required equipment is often cheaper, easier to use or portable. Based on the principle of operation, dedicated methods can be divided into mechanical, optical and acoustic methods.

One mechanical method introduced in [11] measures velar position with a spring wire fixed at a molar at one end and touching the lower side of the velum at the other end. Velar movements bend the spring wire and the bending is transformed into an electrical signal with a strain gauge attached to the wire. Another mechanical measurement method called Velotrace uses a lever that is inserted into the nasal cavity and rests on the upper side of the velum [12]. Both methods allow continual measurements of velar movements but are very inconvenient for the user.

Dedicated optical systems are the Velograph [13] and Nasograph [14]. They exploit the effect that the intensity of light transmitted across the velopharyngeal port correlates with the port size and require a soft plastic tube or optical fibres to be inserted through the nasal cavity down into the pharynx. Also these methods are invasive and inconvenient for the user.

Dedicated acoustic methods are non-invasive. One such method is acoustic rhinometry where the cross-sectional area of the nasal cavity is estimated from a broadband impulse emitted into one nostril [15]. Since the cross-sectional area of the posterior nasal cavity varies with velar height, this method can detect the velar state [16]. Drawbacks of acoustic rhinometry are that it requires about one second per measurement, that simultaneous phonation is not possible and that the device is rather big. An alternative acoustic method estimates velar position indirectly from separate recordings of oral and nasal sound pressure levels during normal speech production [17], but does not work for silent speech.

The goal of the present study was to design a minimally invasive acoustic system for velar state detection like rhinometry, which also allows quasi-continual measurements and simultaneous phonation like the optical and mechanical methods. The basic idea was to evaluate the acoustic response of the nasal cavity to wideband chirp signals in the low ultrasound range. The responses should differ between low and high velar states, because velar height modifies the shape of the resonance chamber, i.e., the nasal cavity. Ahmadi et al. [18] and McLoughlin [19] recently introduced a similar approach using low-frequency ultrasound but for mouth state and voice activity detection.

Our earlier attempts to measure the acoustic response of the nasal cavity with a small sensor attached to one nostril confirmed that velar height is indeed reflected in the acoustic response [20]. However, the previous system only worked for silent speech or very soft phonation, but was seriously distorted by simultaneous normal or loud phonation. Furthermore, it had no mechanism to infer the velar state from the acoustic responses of the nasal cavity. Here we introduce an improved system that works reliably both during normal and silent speech

http://dx.doi.org/10.21437/Interspeech.2016-771

and includes an approach to infer the velar state from the nasal cavity echo.

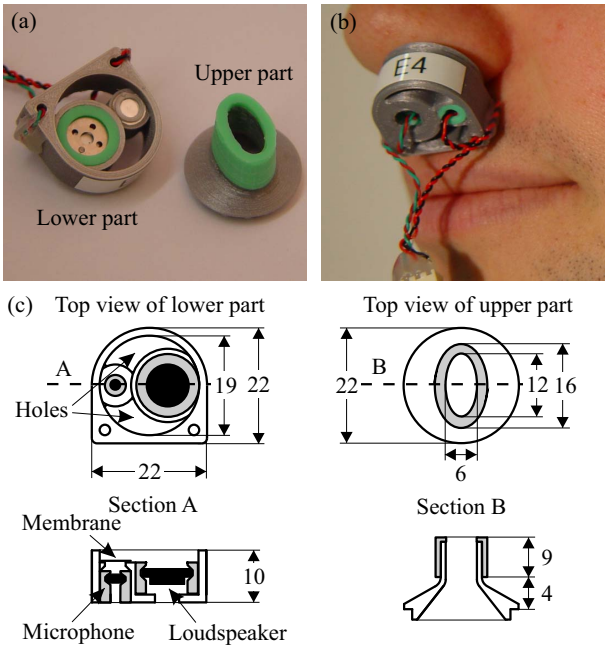

Figure 1: *(a) Lower and upper part of the sensor capsule. (b) Capsule inserted into the nostril. (c) Drawings of the lower and upper parts. The gray parts are made of silicone rubber.*

2. Sensor design

The sensor consists of a 3d-printed capsule (Ultimaker 2 printer with polylactid material) that is divided into a lower and an upper part as shown in Figure 1a and c. The lower part contains a miniature dynamic loudspeaker (type CDM-10008 by CUI Inc.) to excite the nasal cavity with low-frequency ultrasound chirps and an omnidirectional miniature electret condenser microphone (type CME-1538-100LB by CUI Inc.) to record the acoustic response. Loudspeaker and microphone were selected with respect to small size (10 mm and 4 mm diameter, respectively), high upper frequency (20 kHz each according to datasheet) and low cost. To reduce structure-borne transmission of sound from the loudspeaker to the microphone, each of them is held in place by a soft silicone rubber ring (durometer 20 Shore) with a tailored shape. The upper part of the capsule is essentially a funnel to guide the sound from the loudspeaker into the nostril and back to the microphone. The section of the upper part to be inserted into the nostril is covered with a soft silicone rubber skin to prevent injury of the nasal mucosa. Figure 1b shows the two parts of the capsule sticked together and inserted into the nostril.

For the system to work not only during silent speech but also during simultaneous phonation, two conditions must be met:

1. The frequency bands of the chirp signals emitted by the loudspeaker and (nasal) speech sounds must not overlap.
2. The sound pressure level (SPL) entering the microphone must stay below the threshold where the microphone signal distorts or clips.

The first condition is easily met with a lower frequency limit of the chirps of 10 kHz or higher, since the energy of nasal sounds is mainly concentrated below 10 kHz. The second issue is more intricate. Consumer electret condenser microphones with two-pin connections (as the one used in our design) typically start to distort and clip above 100-110 dB SPL[1]. Microphones with a higher maximum SPL are either bigger in size or much more expensive. MEMS microphones often have a somewhat higher maximum SPL of 120 dB, but need specialized circuitry for control. The electret microphone used in our design was found to clip at 109 dB SPL.

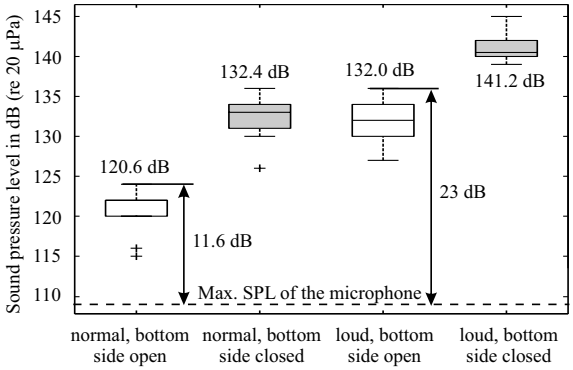

Figure 2: *SPLs measured in the sensor capsules during normal and maximally loud phonation, with and without openings in the bottom side of the capsule. The numbers next to the box plots are the mean values of the corresponding distributions.*

As shown in Figure 2 the SPL in the sensor capsule during the production of a nasal is usually much higher than 109 dB and would immediately cause clipping of the microphone signal. The values in Figure 2 were determined with a measurement microphone (type MK301E with amplifier MV310 by Microtech Gefell GmbH, Germany with a maximum rated SPL of 158 dB) that was inserted into the sensor capsule from the bottom side instead of the loudspeaker. SPLs were measured when five subjects produced /m/ with normal and maximally loud voice with the capsule inserted in either of the two nostrils. The gray boxplots show the SPL distributions for the case that the bottom side of the capsule was completely closed, so that no sound could escape from the respective nostril. The white boxplots show the SPL distributions during normal and loud voice for the case that the capsule was open at the bottom side around the "pots" for the loudspeaker and microphone (the areas labeled as holes in Figure 1c). Hence, the SPL in the capsule is by 10-12 dB lower when the nostril is kept "open" by means of holes compared to completely closed capsules. Considering basic acoustic theory this is not surprising, because a closed nostril causes an antinode of pressure in the capsule. But even with the open capsules, the SPL during "normal" phonation is still about 12 dB above the clipping threshold of the microphone.

As this requires further reduction of the SPL at the microphone, we experimented with different covers in front of the microphone to damp the acoustic signal. Our first attempt was to use a soft silicone rubber layer of 1.5 mm thickness added on top of the rubber ring support for the microphone, which would be a convenient solution in terms of the production process. To evaluate the resulting damping effect, we measured the transfer function (TF) from the input of an external loudspeaker

[1] All given dB values are with respect to 20 μPa reference pressure.

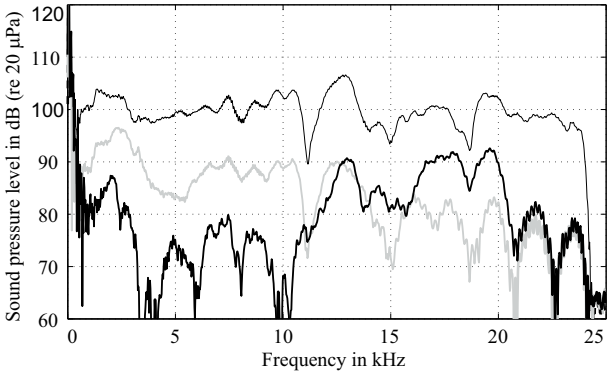

Figure 3: *Transfer functions from the input of an external loudspeaker to the output of the electret condenser microphone in the capsule without any cover in front of the microphone (thin black curve), with a thin rubber layer cover (gray curve) and with a thin membrane as cover (thick black curve).*

to the output of the (covered) microphone between 100 Hz and 24 kHz. Therefore, the bottom part of the capsule containing the electret microphone was mounted at a fixed distance in front of the loudspeaker with the microphone membrane facing the loudspeaker membrane. The TF for the microphone damped with the rubber cover is shown as the light gray curve in Figure 3. In contrast to the accordingly measured TF without any cover in front of the microphone (thin black curve), the rubber cover damps the signal between about 6 dB at low frequencies to about 20 dB at high frequencies. This behaviour is rather undesirable, because in our system we want the low speech frequencies to be damped as much as possible and the high chirp frequencies used for velar state detection to be damped as little as possible. The best way we found to damp mainly the low frequencies was to span a thin membrane slightly above the microphone as shown in Section A in Figure 1c. Such a membrane constitutes an acoustic mass in series with an acoustic compliance that depend on the radius, thickness, density and tension of the membrane [21]. Depending on the parameters, the membrane can act as a band-pass filter. We experimented with foils of different material and thickness and found polyethylene packaging foil with a thickness of 42 μm (as determined with a micrometer screw) to give a satisfactory TF, which is shown as thick black line in Figure 3. Here, speech frequencies up to 10 kHz are damped by at least 20 dB while sound waves pass the membrane with less damping above 10 kHz. Hence, this solution was used for the measurements described below.

3. Sensor operation

Both microphone and loudspeaker in the capsule were connected to a USB audio interface (type AUREON XFIRE8.0 HD by Terratec) that was connected to a standard laptop running the control software. The audio interface was used in full-duplex mode for simultaneous playback and recording at a sampling rate of 96 kHz and 16 bit quantisation. In operation, the loudspeaker in the capsule emitted a sequence of contiguous chirps at a rate of 40 chirps/s, each with a power band from 12-24 kHz and 25 ms duration. The sampling rate of 40 Hz was chosen as a trade-off between the temporal resolution of the velar state detection and the signal-to-noise ratio achievable with a single measurement.

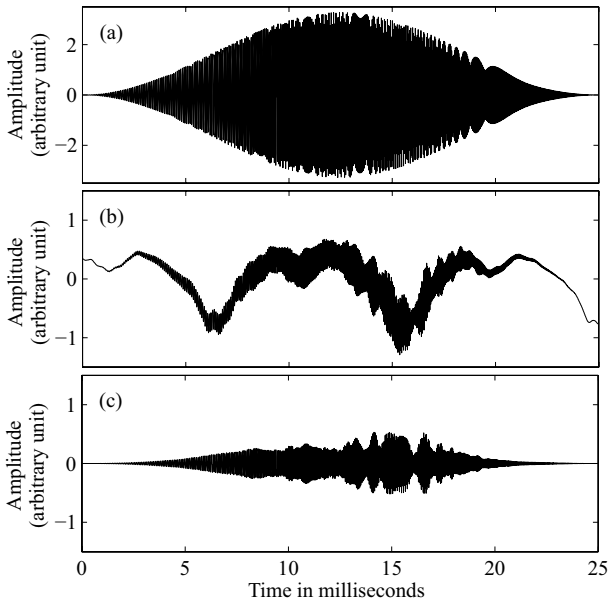

Figure 4: *a) Source signal (chirp). b) Recorded signal during phonation of /m/ (superposition of speech signal and chirp response). c) The chirp response without the low-frequency components from the speech signal.*

The waveform of an individual chirp $s(t)$ was generated analogously to [22] as

$$s(t) = A(t)\sin(2\pi \int_0^t f(\tau)d\tau), \tag{1}$$

where t is the time, $f(\tau)$ is the instantaneous frequency, and $A(t)$ is the temporal envelope of the chirp (a 25 ms long Hanning window in our case). The power band from 12 to 24 kHz starts above the relevant frequencies of speech signals and fully exploits the frequency limits of the loudspeaker and the microphone in the capsule. In order to obtain a constant power density for all frequencies in the power band, the instantaneous frequency had to be varied as

$$f(t) = f_{\min} + (f_{\max} - f_{\min}) \cdot \frac{\int_0^t A^2(\tau)d\tau}{\int_0^T A^2(\tau)d\tau}, \tag{2}$$

where $f_{\min} = 12$ kHz is the lower frequency limit, $f_{\max} = 24$ kHz is the upper frequency limit, and $T = 25$ ms is the chirp length. The last factor on the right hand side compensates the effect of the envelope $A(t)$ on the power density compared to a linear increase of the instantaneous frequency. The chirp waveform is plotted in Figure 4a.

The acoustic response of the nasal cavity to the emitted chirps was continually recorded with the microphone. The recorded signal was high-pass filtered using an 8th order Chebyshev filter with a cutoff frequency of 12 kHz to separate the chirp response from potential low-frequency speech sound. Figure 4b shows an example of the recorded signal during phonation of a nasal, and Figure 4c shows the signal after high-pass filtering, i.e., the pure chirp response. The chirp response was then Fourier transformed and the magnitude spectrum was considered as feature vector for velar state detection. The 25 ms time window for the Fourier transform was automatically centered around the chirp response to exclude signal parts belonging to the previous or the following frame.

4. Test results

Figure 5 shows examples of magnitude spectra of chirp responses obtained during the production of /m/ (solid gray line), during the closure phase of /p/ (solid black line) and during strong inhalation through the nose (dotted line). The (flat) magnitude spectrum of the source chirp is shown as dashed line. It can be seen that the magnitude spectra of the chirp responses differ at multiple places in the 12-24 kHz band between raised and lowered velum states (response for /p/ vs. responses for /m/ and inhalation). In this example, the differences are very pronounced around the dip at 18 kHz. The chirp responses also reflect different degrees of lowering. The response for strong inhalation through the nose (dotted curve) deviates more from the raised velum state (black curve) than the response for the nasal (gray curve). In general we found that the spectral differences are often small and that the frequencies where the differences occur vary across speakers and even between the left and right nostril of the same speaker. On the other hand, the spectral differences are quite consistent in multiple repetitions of sounds with raised and lowered velum states.

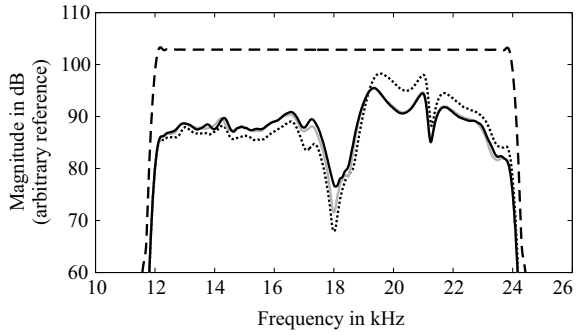

Figure 5: *Magnitude spectra of chirp responses during the production of /m/ (solid gray line), during the closure phase of /p/ (solid black line) and during strong inhalation (dotted line). The magnitude spectrum of the source chirp is shown as dashed line.*

In order to map a chirp response to a single value that indicates the velar state (or height) we devised a measure based on the difference of a given chirp response spectrum $\boldsymbol{X}$ to reference spectra $\boldsymbol{R}$ and $\boldsymbol{L}$ for the raised and lowered velar states, respectively. As distance measure between two spectra $\boldsymbol{X}$ and $\boldsymbol{Y}$ we used

$$d(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{k=k_1}^{k=k_2} (|X_k| - |Y_k|)^2} \qquad (3)$$

where k_1 and k_2 are the frequency indices of the lower and upper power band limits of the chirp and $|X_k|$ is the spectral magnitude at the frequency index k. Based on this distance measure, a "velar height" variable h was defined as

$$h = d(\boldsymbol{X},\boldsymbol{L}) / \left[d(\boldsymbol{X},\boldsymbol{L}) + d(\boldsymbol{X},\boldsymbol{R})\right]. \qquad (4)$$

The variable h varies between 0 and 1 with $h = 0$ corresponding to the lowered velar state and $h = 1$ to the raised velar state.

To test this method, several utterances have been recorded with the sensor from one speaker. At the beginning of each utterance, the speaker produced a sustained /m/ and a sustained /ʃ/, each for about one second. From 20 chirp responses (=0.5 s) in the stationary phase of /m/, an average reference spectrum $\boldsymbol{L}$ for the lowered velar state was calculated. From 20 chirp responses in the stationary phase of /ʃ/, an average reference spectrum $\boldsymbol{R}$ for the raised velar state was calculated. Given theses spectra the velar height h was calculated over the entire utterance and visually analyzed. Figure 6 shows an example of one utterance with the sustained reference sounds at the beginning followed by the nonsense words /ʃmaː/, /ʃnaː/ and /ʃŋaː/. The black curve shows the low-frequency part (< 12 kHz) of the signal recorded with the capsule microphone and the red curve shows the velar height calculated with (4). This example illustrates that the velum height has more extreme values during the sustained reference sounds than during the words. In the words, the velum state is always closest to the raised reference state during /ʃ/, closest to the lowered reference state during the nasals, and inbetween for the vowels. Comparable patterns were observed for the other test utterances.

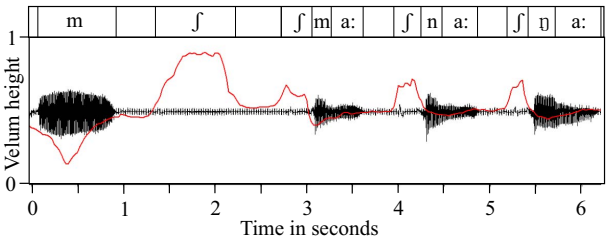

Figure 6: *Calculated velar height (red curve) during a test utterance.*

5. Conclusions and outlook

In this study we presented the design of a novel sensor for velar state detection that is small, lightweight, cheap and works during normal and silent speech. Although the development is still at an early stage, our preliminary tests indicated the feasibility of velar state detection with this kind of sensor. A number of problems remain to be solved. A close examination of chirp response spectra for different speech sounds in different contexts indicated that the response spectrum might vary not only along one dimension (as we implicitly assumed with the two reference spectra for raised and lowered velar states) but also along a second dimension. This would mean that the variation of the nasal cavity shape due to velar movements has not just one degree of freedom (DOF), as often assumed from a functional point of view, but a second DOF, as demonstrated in [23]. Considering variations of the nasal cavity shape and hence the chirp response spectrum along two dimensions would require a new calibration method but could potentially lead to more accurate measurements. However, the calibration method should be kept as simple as possible, because a re-calibration is necessary each time when the capsule is inserted into the nostril. The reason is that the angle and the depth of insertion have an effect on the chirp response spectra. Finally, the reliability of the method could be improved by considering a lower frequency (e.g., 8 kHz) as the lower power band limit of the chirps, because more spectral information would become available to indicate velum state differences. First tests showed that the interference of the chirps with speech would still be acceptably low, but the chirps would become more perceptible and possibly disturbing.

6. Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (BMBF), support code 13GW0101.

7. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] T. Schultz and M. Wand, "Modeling coarticulation in emg-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[3] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *International Joint Conference on Neural Networks*, 2015, pp. 1–7.

[4] T. Hueber, E. L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[5] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech & Language*, vol. 36, pp. 274–293, 2016.

[6] H. J. Künzel, "First application of a biofeedback device for the therapy of velopharyngeal incompetence," *Folia Phoniatrica et Logopaedica*, vol. 34, no. 2, pp. 92–100, 1982.

[7] G. M. Phippen, "A feasibility study of visual feedback speech therapy for nasal speech associated with velopharyngeal dysfunction," Dissertation, University of Southampton, Faculty of Health Sciences, 2013.

[8] K. L. Moll and R. G. Daniloff, "Investigation of the timing of velar movements during speech," *Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 678–684, 1971.

[9] F. Bell-Berti and R. A. Krakow, "Anticipatory velar lowering: A coproduction account," *Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 112–123, 1991.

[10] 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.

[11] K. T. Moller, R. R. Martin, and R. L. Christiansen, "A technique for recording velar movement," *Cleft Palate Journal*, vol. 8, pp. 263–276, 1971.

[12] S. Horiguchi and F. Bell-Berti, "The velotrace: A device for monitoring velar position," *Cleft Palate Journal*, vol. 24, no. 2, pp. 104–111, 1987.

[13] H. J. Künzel, "Röntgenvideographische Evaluierung eines photoelektrischen Verfahrens zur Registrierung der Velumhöhe beim Sprechen," *Folia Phoniatrica et Logopaedica*, vol. 31, no. 3, pp. 153–166, 1979.

[14] R. M. Dalston, "Photodetector assessment of velopharyngeal activity," *Cleft Palate Journal*, vol. 19, no. 1, pp. 1–8, 1982.

[15] O. Hilberg, A. C. Jackson, D. L. Swift, and O. F. Pedersen, "Acoustic rhinometry: evaluation of nasal cavity geometry by acoustic reflection," *Journal of Applied Physiology*, vol. 66, no. 1, pp. 295–303, 1989.

[16] E. J. Seaver, M. P. Karnell, A. Gasparaitis, and J. Corey, "Acoustic rhinometric measurements of changes in velar positioning," *Cleft Palate-Craniofacial Journal*, vol. 32, no. 1, pp. 49–54, 1995.

[17] S. G. Fletcher, I. Sooudi, and S. D. Frost, "Quantitative and graphic analysis of prosthetic treatment for 'nasalance' in speech," *The Journal of Prosthetic Dentistry*, vol. 32, no. 3, pp. 284–291, 1974.

[18] F. Ahmadi, M. Ahmadi, and I. V. McLoughlin, "Human mouth state detection using low frequency ultrasound," in *Interspeech 2013*, Lyon, France, 2013, pp. 1806–1810.

[19] I. V. McLoughlin, "The use of low-frequency ultrasound for voice activity detection," in *Interspeech 2014*, Singapore, 2014, pp. 1553–1557.

[20] P. Birkholz, M. Schutte, S. Preuß, and C. Neuschaefer-Rube, "Towards non-invasive velum state detection during speaking using high-frequency acoustic chirps," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, R. Hoffmann, Ed. TUDPress, Dresden, 2014, pp. 126–133.

[21] U. Marschner and R. Werthschützky, *Aufgaben und Lösungen zur Schaltungsdarstellung und Simulation elektromechanischer Systeme*. Springer Vieweg, 2015.

[22] J. Neumann, *Recording Techniques, Theory and Audiological Application of Otoacoustic Emissions*. BIS-Verlag Oldenburg, 1997.

[23] A. Serrurier and P. Badin, "A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data," *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2335–2355, 2008.