TOWARDS NON-INVASIVE VELUM STATE DETECTION DURING SPEAKING USING HIGH-FREQUENCY ACOUSTIC CHIRPS

Peter Birkholz, Michael Schutte, Simon Preuß, Christiane Neuschaefer-Rube

Clinic of Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen and RWTH Aachen University
peterbirkholz@gmx.de

Abstract: This paper presents our progress towards a convenient and non-invasive real-time method to measure the state of the velum (raised vs. lowered) that works both during normal and silent speaking. The method emits acoustic signals with a power band from 12 to 24 kHz into a nostril and analyzes the "echo" from the nasal cavity. Here we describe two design iterations of the method, present first test results, and outline strategies for further improvements. The results indicate that the method can possibly discriminate not only the raised from the lowered velum state, but also intermediate states. Applications for the method are, for example, basic phonetic research and silent speech interfaces.

1 Introduction

The size of the velopharyngeal opening, which is related to the state of the velum, determines the oral or nasal nature of speech sounds. Measuring the state of the velum is therefore of great interest in phonetic research, e.g., to examine the coordination between velar and lingual gestures, for silent speech interfaces to discriminate nasal and non-nasal sounds, and to study nasality disorders.

The velum state can be observed with either general imaging techniques like radiography or magnetic resonance imaging, or with specialized techniques. The latter are usually cheaper, less complex, and easier to use than the former. Based on the principle of operation, the existing specialized techniques can be divided into optical, mechanical, and acoustic methods. Optical systems are the "Velograph" [7] and the "Nasograph" [3]. They are based on the transmission of light through the velopharyngeal port and require a flexible tube or optical fibers to be placed through the nasal cavity into the pharynx. A mechanical system was presented by Moller et al. [8]. It is based on a displacement transducer made of thin spring wire and a resistance strain gauge that is fixed to the molars and touches the velum at the bottom side to register velar movements. Another mechanical system is the "Velotrace" [6], which transmits velar movements from an internal lever resting on the upper side of the velum to an external lever via rods on the floor of the nasal cavity. While these optical and mechanical systems can continually track velar movements, they are invasive and rather inconvenient for the user.

An non-invasive acoustic method to measure the velum state is "Rhinometry" [5]. This method is based on acoustic reflection and actually provides an estimate of the cross-sectional area of the nasal cavity as a function of distance. As the cross-sectional area in the posterior nasal cavity varies with velar position, this method can successfully detect changes in velar positioning [10]. While this method is non-invasive, it needs about 1 s for a single measurement, it does not work with simultaneous phonation, and the device is rather bulky. Another non-invasive acoustic

method estimates the degree of velopharyngeal opening from separated recordings of the oral and nasal acoustic energies, but works only during normal speech production [4].

In this paper we present our progress towards a new approach to measure the velum state based on acoustic reflections. The main goal was to be able to discriminate whether the velopharyngeal port is open (lowered velum) or closed (raised velum) and possibly to detect different degrees of opening. Our method was supposed to be non-invasive, convenient and cheap, have a sampling rate of at least 40 Hz, and work both with and without simultaneous phonation. Hence, it should combine the ease of use of Rhinometry with the advantages of the optical and mechanical methods (real-time state tracking; simultaneous phonation possible) and finally require only a small and cheap device.

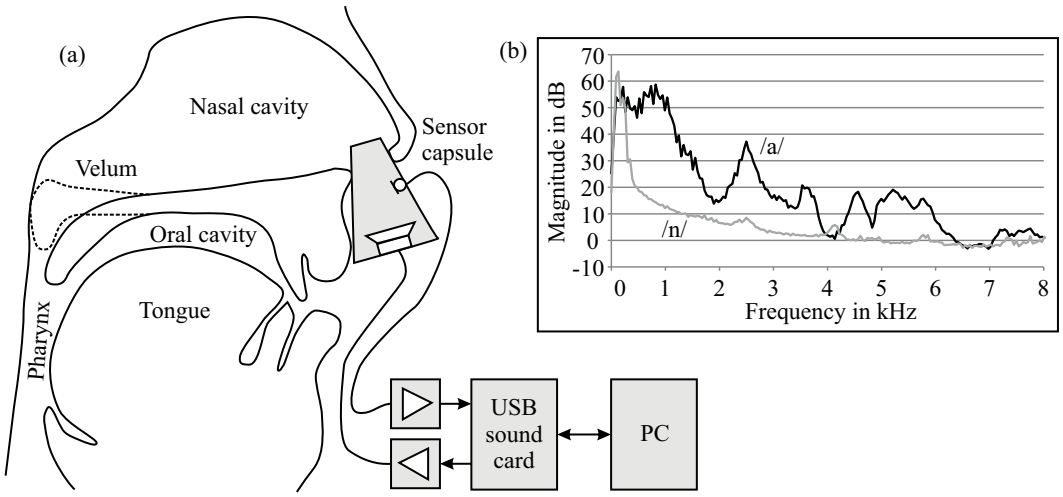

Figure 1 - (a) General setup of the measurement system. (b) Long-term average spectra of /n/ (gray curve) and /a/ (black curve) as measured by a microphone a few centimeters in front of the mouth.

The basic idea was as follows: A frustum-shaped capsule containing a miniature loudspeaker and a miniature microphone is inserted into one nostril as illustrated in Figure 1a. The loudspeaker periodically emits sound bursts with a high-frequency power band into the nasal cavity through the upper open end of the frustum, and the microphone records the sound of the bursts filtered by the nasal cavity. As the resonance characteristics of the nasal cavity change with the degree of velopharyngeal opening, we expected these changes to be reliably detectable in the recorded signals by pattern matching. The high-frequency power band should prevent interference of the measurement with simultaneous phonation. As the long-term average spectra of the sounds /a/ and /n/ exemplify in Figure 1b, speech sounds have only little power left at frequencies of 7 kHz or higher. Somewhat similar to our approach, low-frequency ultrasound, emitted and recorded by a mobile phone a few centimeters in front of the mouth, was recently successfully applied for *mouth* state detection (open vs. partially open vs. closed) [1].

2 First prototype

2.1 System design

The first prototype of our system is shown in Figure 2. It consists of a frustum-shaped capsule with a height of 27 mm and outer diameters of 19 mm and 10 mm at the bottom and top, respectively. The shape of the capsule was designed using the 3D modeling software SketchUp

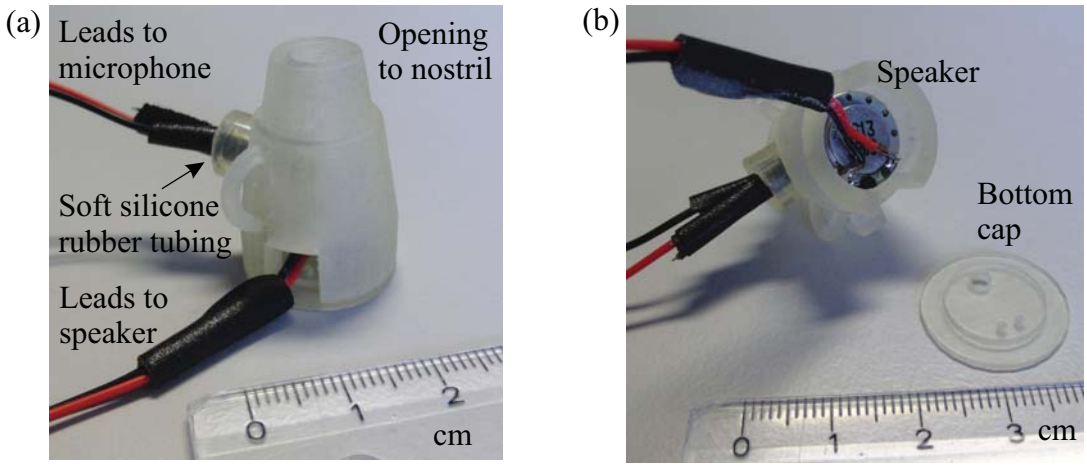

Figure 2 - First design of the capsule in (a) assembled state and (b) disassembled state.

version 8.0 and created with a 3D printer by fused deposition modeling. A dynamic miniature loudspeaker (type KDMG13008C-03 by Kingstate Electronics Corp., 13 mm diameter, 3 mm height, 0.3 W rated power) was flush mounted in the bottom end of the capsule. The upper end of the capsule is open to feed the loudspeaker sound into the nostril. At half the height of the capsule, an omnidirectional miniature electret microphone (type CME-1538-100LB by CUI Inc., 4 mm diameter, 1.5 mm height, 120 dB max. SPL, 58 dBA SNR) was inserted through a hole into the capsule flush with the inner side of the wall. The microphone was embedded in soft silicone rubber tubing to reduce structure-borne sound from the loudspeaker. The upper frequency limits of both the loudspeaker and the microphone were specified as 20 kHz.

The loudspeaker and microphone were connected to a laptop via a consumer USB sound card (type SC010 by SWEEX). This sound card allows simultaneous playback and recording with a sampling rate of 96 kHz and 16 bit quantization. For a high signal-to-noise ratio of the measurements and little harmonic distortions, the loudspeaker should be operated around its rated power of 300 mW. However, the internal amplifier of the sound card was only able to deliver about 140 mW to the 8 Ω loudspeaker, which we estimated from the measured peak amplitude of the output voltage $\hat{V} = 1.5$ V with the equation $P = \hat{V}^2/2R_{\text{speaker}}$. Therefore, we connected a custom-built USB-powered audio amplifier based on the bridge-connected audio power amplifier LM4861 (Texas Instruments) between the sound card output and the loudspeaker. The microphone did not require external amplification and was directly connected to the sound card. The sensitivity of the microphone was in fact so high that we had to set the recording level to about 5% to avoid clipping during the measurements.

A customized program running on the laptop was created for playback and recording control, data processing, and visualization. Audio data were processed with 96 kHz/16 bit. The source signal used to drive the loudspeaker was a band-pass filtered unit impulse with a passband (i.e., power band) between 12 and 24 kHz. The unit impulse was filtered with a high-pass and a low-pass Chebyshev filter (8th order each) and then normalized in amplitude. Figure 3a shows the waveform of the source signal, and the upper black curve in Figure 3b shows its magnitude spectrum. The source signal was emitted at a rate of 40 Hz in order to allow quasi-continuous velum state tracking. Given that raising and lowering gestures of the velum have a minimum duration of 100 ms [11], 40 Hz appears to be a sufficiently high rate for velum state sampling.

To perform a measurement, the capsule has to be pushed into one nostril as illustrated in Figure 1a. Then the source pulses are emitted by the loudspeaker, passing through the capsule into the nasal cavity. The sound pressure in the capsule is simultaneously recorded with the mi-

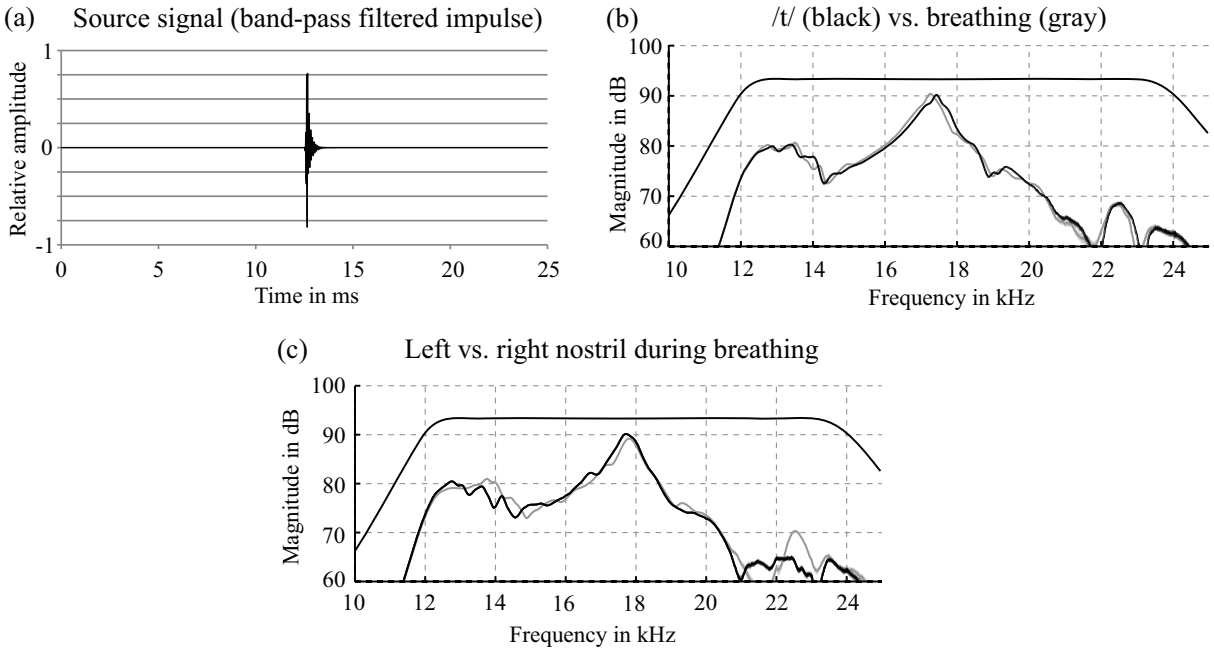

Figure 3 - (a) Source signal. (b) Spectra for the closure phase of /t/ vs. breathing (low velum) measured for the left nostril. (c) Spectra for a low velum during breathing measured at the left and right nostrils. The shaded regions indicate the confidence interval from the mean ± 1 standard deviation. They are generally very small and visible for low magnitudes only.

crophone. After measurement, the recorded signal is high-pass filtered with a cutoff frequency of 12 kHz using an 8th order Chebyshev filter to eliminate a potential voice signal. Hence, the filtered signal contains essentially the response of the nasal cavity to the emitted source pulses. For each recorded pulse, the magnitude spectrum is calculated, because these spectra are likely to reflect differences in velum state. Therefore, each pulse is multiplied with a custom window and transformed into the frequency domain by FFT. The custom window is centered around the pulse and has the form

$$w_i = \begin{cases} 0.5(1-cos(\pi i/K)) & \text{if } 0 \leq i < K \\ 1 & \text{if } K \leq i < M-K \\ 0.5(1+cos(\pi(i-M+K)/K)) & \text{if } M-K \leq i < M \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $M = 2400$ is the window length in samples (corresponding to the frame length of 1/40 s at the audio sampling rate of 96 kHz) and $K = 240$ is 1/10th of the window length. Hence, the window is essentially rectangular with a smooth decay towards zero at the ends. This flat-top type of window was chosen to preserve the shape of the recorded pulse in the center of the window, but to prevent spectral distortions if the signal in the high-frequency band happened to be not completely zero in the regions *between* the pulses, e.g., due to high-frequency noise. To consistently center the windows for calculating the spectra, the precise position of the recorded pulses is detected using cross-correlation between the source pulses and the recorded pulses. The lag between the two is apparent as a clear peak in the correlation function.

2.2 Tests

To test the system, we recorded and compared pulses for raised and lowered velum states. For the first test, one subject with the capsule in place produced a silent /t/ by closing the vocal tract with the tongue tip, raising the velum, and generating an excess pressure for about 2 s. Then the velum was lowered to breath out through the nose for another two seconds. From the stationary portions in the middle of both phases the spectra of about 20 recorded pulses were obtained. Figure 3b shows the average spectra for both conditions. The variation of the spectra over the 20 pulses is shown by shaded regions around the average spectra extending from the mean ± 1 S.D. For both conditions, the variation was so small that it can hardly be seen. As Figure 3b shows, there were clear spectral differences between the raised and lowered velum states, especially in the regions around 14 and 17 kHz.

In a second test, we examined the difference between the left and right nostril for the lowered velum state (normal breathing). Figure 3c shows that there were great spectral differences between the nostrils, and that they were greater than the differences between the raised and lowered velum state measured at one nostril (Figure 3b). This indicates that the system needs to be calibrated individually for each nostril.

In a third test, the subject produced each of the phones /n/ and /s/ for multiple seconds to contrast the raised and lowered velum states during normal speech production. While we obtained good spectra for /s/ (similar to the silent /t/), we encountered serious problems with /n/, because the sound pressure level (SPL) in the capsule became very high due to phonation. This high pressure level had two negative effects. Firstly, the recording level had to be strongly reduced to avoid clipping of the signal. Lowering the recording level also reduces the intensity of the high-frequency pulses in the signal and hence the SNR of the measurements. Secondly, and even worse, the recorded pulses varied strongly and apparently unpredictably (despite the absence of clipping) so that the signals were deemed unusable. The second effect may have two causes, namely that the SPL in the capsule exceeded the maximum rated SPL of the microphone, resulting in strong harmonic distortions, and that the high SPL alters the normal vibration of the loudspeaker membrane in a non-linear way.

3 Second prototype

3.1 System design

The main limitation of the first prototype was the very high SPL in the capsule during production of nasal sounds. The reason for this was most probably that the capsule completely closed the nasal cavity at one nostril, because it is known from acoustic theory that standing waves in tubes reach pressure maxima at closed ends of tubes. Hence, for the second prototype, the capsule design was altered to have an opening at the bottom end.

The new design is shown in Figure 4. It has again the shape of a frustum, but with slightly greater dimensions (30 mm height, outer diameters of 22 mm and 9 mm at the bottom and top, respectively). The microphone is of the same type and mounted in the same way as in the first design. However, the loudspeaker is somewhat smaller (type KDMG10008C-03 by Kingstate Electronics Corp., 10 mm diameter, 3 mm height, 0.3 W rated power) and embedded in a rubber ring on the bottom cap to further reduce structure-borne sound transmission from loudspeaker to microphone. Next to the loudspeaker, a banana-shaped hole was cut into the bottom cap to reduce the occurrence of high acoustic pressures in the capsule during nasal sound production.

With respect to the source pulse, the filtered impulse of the first design was replaced with a

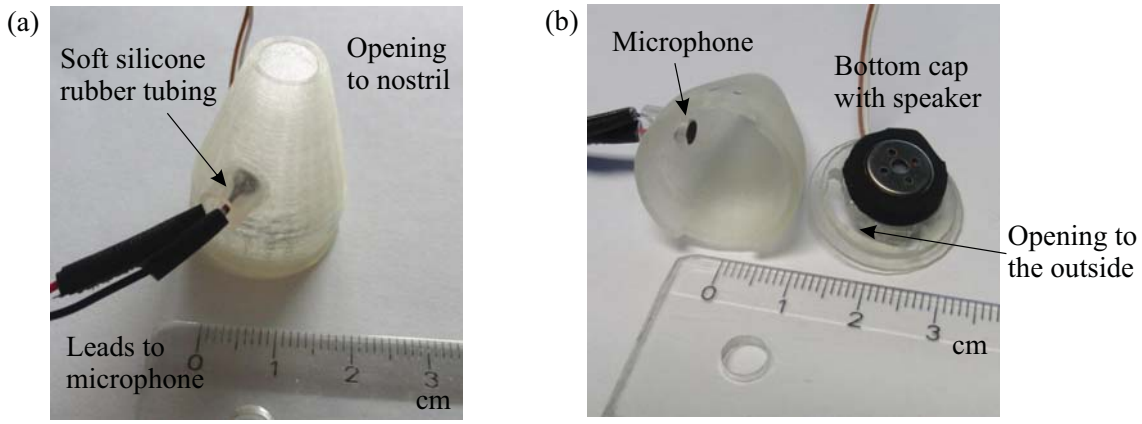

Figure 4 - Second design of the capsule in (a) assembled state and (b) disassembled state.

chirp between 12 and 24 kHz. Compared to impulses, chirps have the lowest possible peak to RMS amplitude ratio and can achieve the highest feasible signal to noise level [2]. In fact, the energy in the chirp signals was 88.4 times the energy in the filtered impulses used in the first design. This promised to increase the SNR of the measurements.

The actual waveform of the chirps was generated analogously to [9] as

$$s(t) = A(t)\sin(\phi(t)), \tag{2}$$

where $A(t)$ is the envelope of the chirp (a 25 ms long Hanning window in our case) and

$$\phi(t) = 2\pi \int_0^t f(\tau)d\tau \tag{3}$$

is the phase. $f(\cdot)$ denotes the instantaneous frequency, which would linearly increase from the lower to the upper frequency limit for linear chirps. However, with a linear increase of instantaneous frequency, the signal power between the lower and upper frequency limits would vary according to the envelope $A(t)$. To obtain constant power for all frequencies in the power band, the instantaneous frequency must increase more slowly for lower envelope amplitudes. This was achieved with the equation

$$f(t) = f_{\min} + (f_{\max} - f_{\min}) \cdot \frac{\int_0^t A^2(\tau)d\tau}{\int_0^T A^2(\tau)d\tau}, \tag{4}$$

where $T = 25$ ms is the window length. The power band limits were set to $f_{\min} = 12$ kHz and $f_{\max} = 24$ kHz as in the first design. The new source signal is shown in Figure 5a and the corresponding magnitude spectrum is the upper black curve in Figure 5b. Apart from the capsule design and the source pulse shape, the other aspects of the measurement system are the same as for the first prototype.

3.2 Tests

In the first step, we compared the spectra of the recorded pulses (chirps) for a silent /t/ (raised velum) and during breathing (lowered velum) again. Figure 5c illustrates that there were pronounced spectral differences between these velum states at multiple frequencies marked by the arrows. Note that these spectra differ from Figure 3b (first design) due to the new loudspeaker type and capsule shape. Next, we obtained averaged spectra for sustained /s/ (raised velum) and /n/ (lowered velum) produced at a normal loudness level. This time, the SPL in the capsule

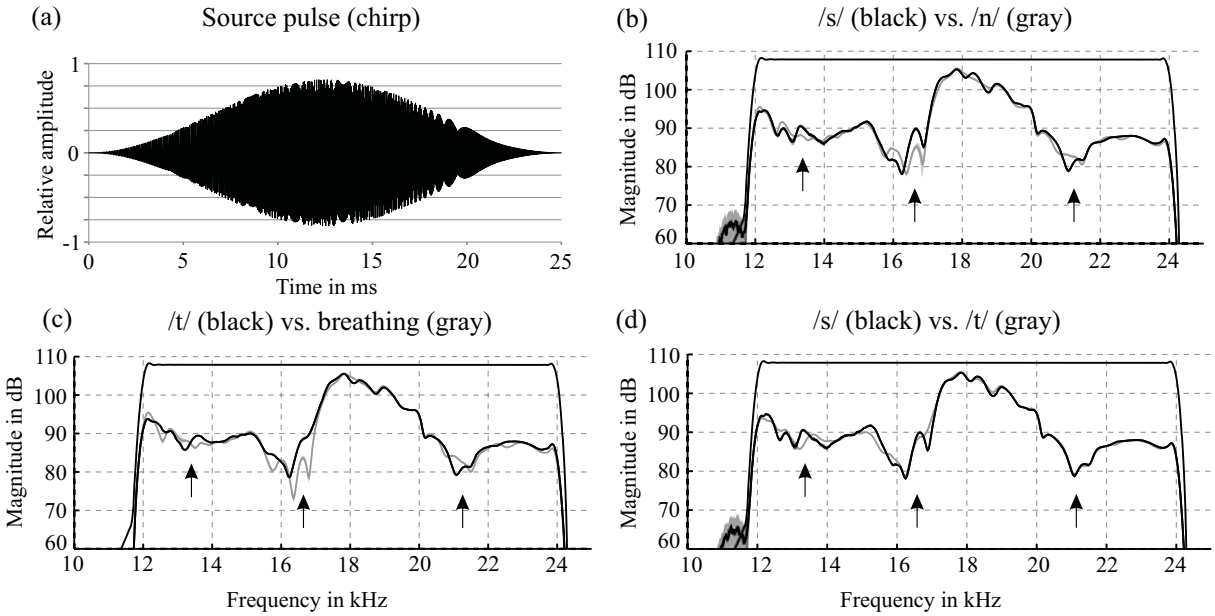

Figure 5 - (a) Chirp signal. (b) Spectra for /s/ vs. /n/. (c) Spectra for the closing phase during /t/ vs. normal breathing. The shaded regions indicate the confidence interval from the mean ± 1 standard deviation. They are generally very small and visible for low magnitudes only.

was sufficiently low to greatly reduce the negative effects encountered with the first prototype. Figure 5b shows that the differences between the raised and lowered velum states were very similar to the differences during the silent articulations shown in Figure 5c (black arrows are at the same frequencies). Apparently, the opening in the bottom cap of the capsule helped to reduce the SPL during nasal sound production sufficiently to prevent excessive harmonic distortions of the microphone and negative effects on the loudspeaker membrane, and hence to obtain stable measurements. Additional measurements showed that the average signal amplitude of the recordings with an open capsule decreased to about 1/4 ($= -12$ dB) of the amplitude with a closed capsule (as in the first design). Interestingly, the opening of the capsule mainly reduced the acoustic pressure in the low-frequency band ($0 - 12$ kHz), i.e., the sound pressure caused by the voice, while the amplitude of the recorded high-frequency pulses was almost the same for the open and closed capsules. This is probably caused by the difference in the diffraction of low and high frequency sound waves. Finally, we compared the spectra for /s/ and /t/, which are shown in Figure 5d. Although the velum is raised in both conditions, there are still minor differences in the recorded pulse spectra (which are much smaller than differences between raised and lowered velum states). A closer analysis of the *transitions* between lowered and raised velum states (for example in the consonant cluster /nt/) showed that these differences can be well explained by the fact that the actual velum position during /s/ is somewhat lower than during the closing phase of /t/.

4 Outlook

This study presented a working prototype of a system to detect differences in velum state using high-frequency pulse reflectometry. The advantages of the system are that it is non-invasive and convenient, allows real-time tracking at 40 Hz, works with and without simultaneous phonation, and is conceptually simple and cheap. However, a few issues remain to be addressed before it

can be considered fully matured. Firstly, we presented only qualitative measurement results in terms of pulse spectra. For automatic velum state detection, the system must be extended with methods for pattern recognition and calibration. Secondly, for very loud productions of nasal sounds, the problems with the high SPL in the capsule can surface also in the second design and distort the measurements. Here, further types of miniature loudspeakers and new capsule designs must be examined. Finally, the part of the capsule that is inserted into the nostril can possibly be improved for a better fit to different nostril shapes. After all, we are confident that this approach will be highly effective for phonetic research and for velum state detection in silent speech interfaces.

5 Acknowledgments

We would like to thank Dietmar Faßbänder for 3D printing the capsules, Ian McLoughlin and Gottfried Behler for useful discussions, and Ingmar Steiner for proofreading the paper.

References

[1] F. Ahmadi, M. Ahmadi, and I. McLoughlin. Human mouth state detection using low frequency ultrasound. In *Interspeech 2013*, pages 1806–1810, Lyon, France, 2013.

[2] J. C. Burgess. Chirp design for acoustical system identification. *Journal of the Acoustical Society of America*, 91(3):1525–1530, 1992.

[3] R. M. Dalston. Photodetector assessment of velopharyngeal activity. *Cleft Palate Journal*, 19(1):1–8, 1982.

[4] 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*, 32(3):284–291, 1974.

[5] 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*, 66(1):295–303, 1989.

[6] S. Horiguchi and F. Bell-Berti. The velotrace: A device for monitoring velar position. *Cleft Palate Journal*, 24(2):104–111, 1987.

[7] H. J. Künzel. Röntgenvideographische Evaluierung eines photoelektrischen Verfahrens zur Registrierung der Velumhöhe beim Sprechen. *Folia Phoniatrica et Logopaedica*, 31(3):153–166, 1979.

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

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

[10] E. J. Seaver, M. P. Karnell, A. Gasparaitis, and J. Corey. Acoustic rhinometric measurements of changes in velar positioning. *Cleft Palate-Craniofacial Journal*, 32(1):49–54, 1995.

[11] K. N. Stevens. *Acoustic Phonetics*. The MIT Press, Cambridge, Massachusetts, 1998.