A HEAD-MOUNTED CAMERA SYSTEM FOR THE MEASUREMENT OF LIP PROTRUSION AND OPENING DURING SPEECH PRODUCTION

Fabian Klause, Simon Stone, Peter Birkholz

Insitute of Acoustics and Speech Communication, Technische Universität Dresden
fabian.klause@mailbox.tu-dresden.de

Abstract: In this paper, we present a head-mounted lip observing camera system and the downstream processing steps and tools needed to obtain protrusion and opening of a subject's lips during speech production. The helmet-like camera system consists of two consumer market cameras mounted directly to a subject's head to keep their positions constant with respect to the skull, thus no post processing image stabilization or feature tracking is needed. The mounting system is built from aluminum pipe segments and 3D-printed hinges and joints. The cameras capture the profile and frontal view of the subject's lips. In the two obtained image sequences, the lip contours are detected using the segmentation software *GlottalImageExplorer*. In these two 1-Bit lip contour images, several extrema are detected and converted to the desired parameters protrusion and opening. Exemplary results are presented for selected vowel-consonant-vowel sequences. All necessary files for the construction of the camera helmet and the extraction of the parameters are provided free of charge at `http://vocaltractlab.de/index.php?page=birkholz-supplements`. The video capturing and post-processing exclusively uses a chain of freeware tools to make the system available to the speech research community.

1 Introduction

The lips are the link between a speaker's vocal tract and the surrounding air. Consequently, they are directly involved in the production of speech sounds. There are mainly two degrees of freedom that characterize the lips' movements: protrusion and opening [1]. The lip protrusion is primarily significant for the articulation of vowels, as it extends or shortens the vocal tract and in this way changes its transfer function resulting in shifting formant frequencies. A change in the degree of lip opening results in a size variation of the sound emitting area and thus in a change of the acoustic impedance, which has an impact on the transfer function of the vocal tract [2]. A determination of these lip parameters is therefore of interest in phonetic and articulatory research, but also in various other fields, e.g., for automatic lip reading to support acoustic speech recognition systems[3, 4] or as a biofeedback for language or pronunciation training. However, capturing these parameters in fluent speech is difficult to accomplish outside of a purely laboratory setting. In [4], the lips' movement is captured by tracking fixed points in the profile view of a subject's face. In the first step, the subject's face is identified followed by the tracking of several facial features. A similar procedure to our approach was developed by Kumar *et al.* in [3], where profile and frontal views of the lips are used to extract lip parameters. In both views a brightness threshold is used to identify the border between lips and background respectively lips and darker interior of the mouth. However, Kumar *et al.* used a threshold on the red channel of the image and thus extracted the *outer* contour of the lips, which depends on the subject's lip width as well as the lip opening. In addition to that, the video capturing for the study presented in [3] was done in a studio setting but it remains unclear how the relative positions of the cameras to the subject were kept constant. In this paper, we build on the work in [3] to develop a more robust lip parameter extraction method and introduce a head-mounted camera fixture to improve the intra-individual consistency of the recorded image sequences. All necessary files to produce the camera helmet and editable example scripts for the extraction of the parameters are made publicly available at `http://vocaltractlab.de/index.php?page=birkholz-supplements`.

2 The head-mounted camera system

The camera system has to fulfill the following requirements: It should be as light and small as possible to ensure high portability and flexible use even outside of a typical test environment. Furthermore, the cameras should stay in absolute constant position and orientation with respect to the lips during a recording session. Lastly, the system has to be easily adjustable for varying anatomy between different subjects. With those constraints in mind, we developed the system that is shown in Figure 1. Instead of suppressing the subject's head movements by a clamping mechanism (as in [5]), the cameras are attached to their head using a helmet-like contraption of adjustable straps. Two mounting plates are located above the subject's temples. On the left plate, the camera fixture, consisting of aluminum rods and 3D-printed joints is mounted. Two LED-Arrays are attached to the system to illuminate the lip area for maximum contrast. The right mounting plate holds a black cardboard sheet that serves as a background for the profile view camera to ensure high contrast between lip-tissue and background. The cameras (*Creative® LIVE! CAM Sync HD*), capture with a resolution of 256x256 pixels at a frame rate 30 fps. According to [6, 46], the lips perform a maximum of 5.7 to 7.7 changes in configuration per second during speech production. A frame rate of 30Hz should therefore be enough to capture every single lip configuration in at least 4 to 5 Frames. The recorded video data is stored as Audio Video Interleaved (AVI) files.

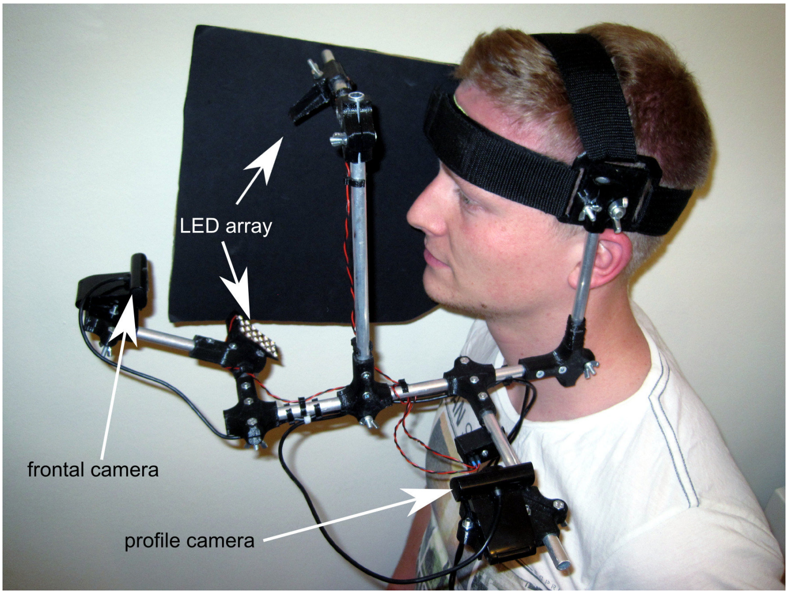

Figure 1 – Head mounted camera system attached to a subjects head

3 Lip parameter definition

The lip observation system provides video sequences of the profile and frontal view of the subject's lips. In these two views, the protrusion and opening (respectively) of the lips can be identified relatively easy: We defined the opening to be the distance between the center of the upper and the lower lip in the frontal view (see [7]). To determine this distance, we drop the perpendicular in the middle of the connection line between both corners of the mouth. The distance of the two points where this perpendicular cuts the upper/lower lip contour, is defined as the absolute opening. To obtain the relative opening, this absolute value is divided by the largest occurring opening in a set of given measurements. The value for the lip protrusion is represented by the horizontal distance between the foremost point of the upper lip in the profile view and a reference plane parallel to the frontal-plane (see [7]), standardized to the interval between its largest and smallest occurring value in a given series of measurements. Since the camera mounting system ensures stable conditions, the left edge of the profile view images serves as the reference plane for the calculation of the protrusion.

The standardization of the measured parameters is not strictly necessary but advantageous if relative changes across several subjects are of interest. It can be replaced by the absolute values if only a single subject is of interest. The significant points for the calculation of the lip parameters in the frontal and profile view can be seen in Figure 2.

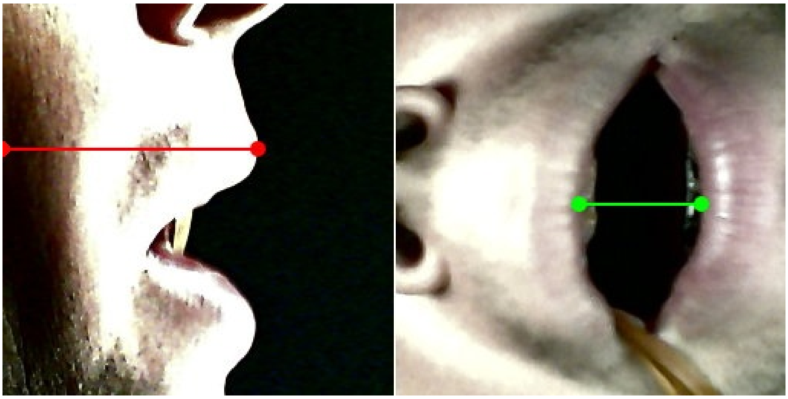

Figure 2 – Definition of the critical points to determine the lip parameters protrusion (red) and opening (green) in the profile (left) and frontal (right) view of a speaker's lips. The vertical lines represent the absolute measured distances between the left border of the image and the upper lip tip and between the upper and lower lip. The frontal view is rotated by 90° to improve the contour tracking in *GlottalImageExplorer*.

4 Processing Steps

After a set of video sequences has been recorded, the video files are further processed in a chain of freeware software tools to extract the relative protrusion and opening parameters. First, the video processing utility *VirtualDub* (http://www.virtualdub.org/) is used to crop the video data to the region of interest and enhance the brightness and contrast of the video data. The processed video data is then opened with *GlottalImageExplorer* (*GIE*) [8]. This software was originally developed to segment the glottis area in endoscopic videos, but the implemented algorithm based on a brightness threshold to detect the border between a dark background bounded by brighter tissue can be straight-forwardly applied to images of the lips and the background in the profile view, respectively the lips and the mouth opening in the frontal view. This yields 1-Bit contour images of the lip profile and the mouth opening. These contours are then evaluated using an *Octave* [9] script to calculate the desired parameters protrusion and opening.

4.1 Video capturing

Before video data can be recorded, the camera fixture has to be firmly attached to the subject's head. Due to the adjustable joints, the cameras can be fixed in any position and orientation with respect to the lips. The frontal camera should come to rest at the same height as the mouth in the sagittal plane, at a distance that allows unobstructed articulation (e.g., maximum spreading [i], maximum opening [a]) and oriented so that both corners of the mouth appear level in the frames. In the profile view, the foremost point of the upper lip has to be captured and thus angular displacement of the profile camera leading to distortion of the face profile has to be avoided. The orientation should be set so that the line between the upper and the lower lips' tips in a relaxed opened position of the mouth stands exactly vertical in the middle of the picture. This ensures that the direction of protrusion movements mainly consists of a horizontal component. The software used for the segmentation, *GIE*, requires the contrast between lips and mouth interior to be as high as possible. Because the speaker's teeth are usually brighter than the lip tissue they may adversely influence the contour tracking. To solve this problem, we used theater-grade black tooth paint to increase the contrast between lip tissue and teeth.

As already mentioned above, video capturing is done with 30 fps. We use a lossless AVI video codec to capture and crop the input video stream to a format of 256x256 Pixels. Assuming a range of motion of the lips of half the image size leads to approximately 130 quantization steps for the captured parameters. As it turned out, this assumption is true in the frontal view for lip opening. The profile view only provides 30-60 steps of quantization depending on the speaker for the lip protrusion, due to the smaller absolute motion range.

In our proposed setup, the two video streams are captured separately in two different instances of *VirtualDub*, launched one after another by a batch script, and thus have to be synchronized after capturing. Because we determined the relative shift to be only a maximum of 0.1 frames per minute, the synchronization can be done by a linear time-shift of one video with respect to the other. To generate a synchronization event at the beginning of a test series, the subject over-articulates the utterance [pa]. The burst onset of the phone [p] can be easily detected in both the front and profile view and the two video streams are manually aligned accordingly by deleting all images prior to the synchronization event. The front video files are converted into a bitmap image sequence after performing white balancing, gamma correction and levels transformation in *VirtualDub* to improve downstream lip contour detection. These processing steps are automated by another batch script.

The segmentation of the lip contours in the bitmap image sequences is done in *GIE* for both the frontal and the profile view. *GIE* was originally intended to segment the glottis area, which has a mostly vertical orientation. The mouth opening is however mostly horizontally oriented. Therefore, each individual image in the frontal image sequence is rotated by 90° in *VirtualDub* before further processing it in *GIE*.

4.2 Frontal image processing

The lip opening can be observed in the frontal view of the lips.

GIE determines and tracks the contour in a image sequence of a relatively dark opening surrounded by relatively bright tissue. It does this using the seeded region growing algorithm. Starting with an initial set of three seed points (pixels) in the area of the mouth respectively the background (in the profile images), the algorithm evaluates the intensity of the neighboring pixels and adds them to the segmented area if their intensity is below a certain threshold. This process is iterated until no more neighboring pixels satisfy the threshold criterion. The recovered 1-bit contours of the left and right border (two contours per input frame), representing the upper and lower lip, are exported in a space-separated text (.txt) file. Each image is represented by two lines in the text file (left and right contour), 256 entries each, containing the horizontal position (in pixels) of left and right border per image row. If no border is detected, the corresponding entry in the text file is filled with a zero. This data file is processed by an *Octave* script to find the center points of the upper and lower lips' contours and calculate the time-varying change of the relative lip opening by sequentially processing each line as outlined in section 3. Figure 3 shows an example frame from the video data and the corresponding contours after lip parameter calculation during the articulation of [u].

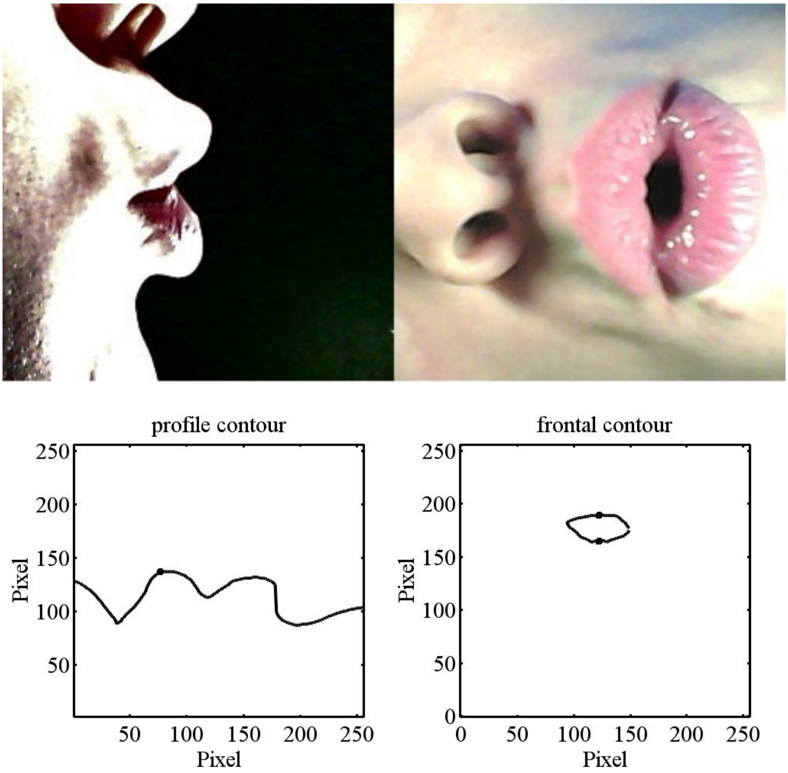

Figure 3 – Captured frontal (right) and profile (left) view of a speaker articulating the vowel [u] after image processing (top row) and their calculated contours with highlighted points for the parameter calculation. The profile contour is rotated by 90° to use the *Octave* function *findpeaks* to find the extreme points of the profile contour. The frontal images are rotated too, to keep the array formats of frontal and profile contours identical throughout the calculations.

4.3 Profile image processing

The lip protrusion can be extracted from the profile view of the speaker's lips. Just as with the frontal view data, the profile view image sequence is imported into *GIE*. In contrast to the frontal images, the area to segment by the program is essentially less „glottis-like". Nevertheless, segmentation is in fact a lot simpler, due to the uniformly black background. The seed points

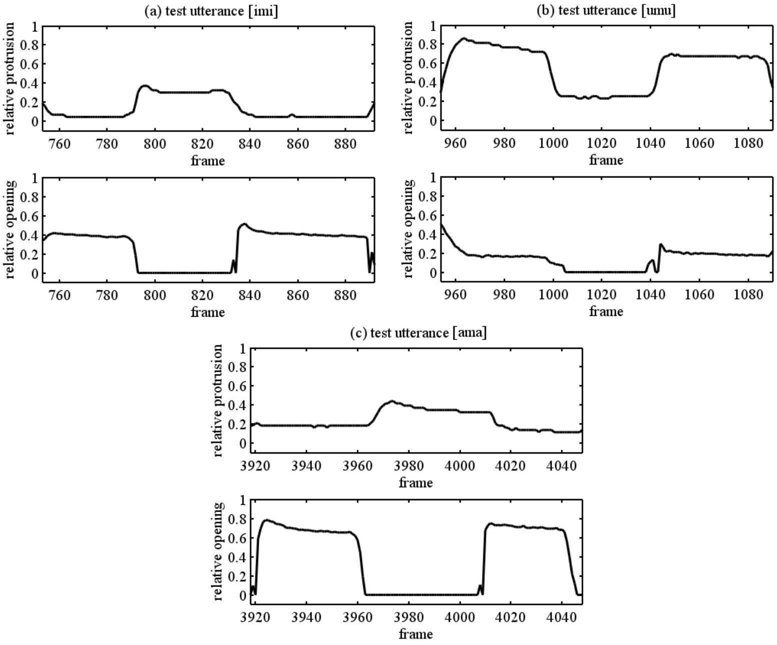

Figure 4 – Example time series of the lip parameters of a subject articulating the test utterances: (a) [imi], (b) [umu] and (c) [ama]

are placed evenly spaced along a vertical axis in the black area and the thresholds are adjusted until the left contour of the segmented area matches the face profile. The right contour ends up at the right border of the image and is irrelevant for further calculations. The exported face contour is then processed by the *Octave* script. It calculates the highest points of the face contour. The first maximum represents the bottom side of the nose, the second one corresponds to the tip of the upper lip and is used to calculate the relative protrusion as outlined above.

5 Results

So far we carried out a test series determining the lip parameters of four speakers articulating test utterances of the character V-[m]-V, where V stands for the vowels: [a e i o u ɛ ø y]. Figure 4 shows example time series of the lip parameters of a subject articulating the test utterances: (a) [imi], (b) [umu] and (c) [ama]. They all show the expected behavior: Regarding the test utterance [umu], we would expect high protrusion during the articulation of the [u] and medium protrusion during the m. The opening is expected to stay close to zero during the entire utterance. In contrast to that, articulating the utterance [ama] should lead to small lip protrusion changes over time, whereas the opening should change from high to low and back to high again during the articulation of the utterance.

6 Summary and Outlook

The proposed recording setup and post-processing methods are evidently suitable for the recording and extraction of the two lip parameters lip opening and lip protrusion. The system is portable (if a laptop is used for the video recordings), easy and cheap to manufacture and assemble, and uses only freely available software. We encourage all members of the speech research community to use the template files for 3D printing of the components and the post-processing scripts made available at `http://vocaltractlab.de/index.php?page=birkholz-supplements` in combination with this paper and the freeware toolchain described above to devise and perform their own studies on labial articulation. While the exemplary results shown above are plausible, there is currently a lack of a reference system to determine the absolute accuracy of our setup. Future studies should be conducted to evaluate this. For example, a human speech scientist could segment a database of test utterances manually and the resulting lip contours could be compared to the semi-automatically generated ones. Another current limitation of our system is that it is not capable of real-time extraction of the lip parameters. However, the proposed methods can straight-forwardly be extended to an on-line real-time processing pipeline. The reference values for maximum and minimum opening and protrusion would have to be recorded separately but once this calibration is done, the algorithms are capable to run in real-time. Another possible improvement would be the inclusion of synchronized audio recordings, e.g., to enable the accurate investigation of transient articulatory movements. Since the cameras used in this setup already feature internal microphones, this can be easily integrated into the existing workflow.

References

[1] GRASSEGGER, H.: *Phonetik, Phonologie*. Schulz-Kirchner Verlag, Idstein, Germany, 5th edn., 2016.

[2] FANT, G.: *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Description and Analysis of Contemporary Standard Russian. De Gruyter, 1970.

[3] KUMAR, K., T. CHEN, and R. M. STERN: *Profile view lip reading*. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–429. IEEE, 2007.

[4] LUCEY, P. and G. POTAMIANOS: *Lipreading using profile versus frontal views*. In *Multimedia Signal Processing, 2006 IEEE 8th Workshop on*, pp. 24–28. IEEE, 2006.

[5] STONE, M. and E. P. DAVIS: *A head and transducer support system for making ultrasound images of tongue/jaw movement*. *The Journal of The Acoustical Society of America*, 98(6), pp. 3107–3112, 1995.

[6] POMPINO-MARSCHALL, B.: *Einführung in die Phonetik*. Walter de Gruyter, 2 edn., 2003.

[7] ABRY, C. and L.-J. BOË: *"Laws" for lips*. *Speech communication*, 5(1), pp. 97–104, 1986.

[8] BIRKHOLZ, P.: *GlottalImageExplorer–An open source tool for glottis segmentation in endoscopic high-speed videos of the vocal folds*. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, 2016. (O. Jokisch, ed.), (Dresden), TUDPress.

[9] EATON, J. W., D. BATEMAN, S. BATEMAN, and R. WEHBRING: *GNU Octave version 4.2.0 manual: a high-level interactive language for numerical computations*, 2016. URL `http://www.gnu.org/software/octave/doc/interpreter`.