Semi-Automatic Processing of Real-time MR Image Sequences for Speech Production Studies

Erik Bresch[1], Jason Adams[1], Arthur Pouzet[1], Sungbok Lee[1,2], Dani Byrd[2], Shrikanth Narayanan[1,2]

[1]Speech Analysis and Interpretation Laboratory, USC Viterbi School of Engineering, Los Angeles, CA, USA

[2]USC Department of Linguistics, Los Angeles, CA, USA

bresch@usc.edu, jason.adams@alumni.usc.edu, pouzet@usc.edu, sungbokl@usc.edu, dbyrd@usc.edu, shri@sipi.usc.edu

Abstract. *This paper describes algorithms and data processing procedures developed to process real-time magnetic resonance images for speech production studies. These techniques allow visualization of the teeth, contour tracking over time of active and passive articulators, and vocal tract aperture recovery. These techniques can facilitate speech production studies requiring, for example, the computation of the resonance frequencies of the moving vocal tract or an examination of intergestural timing. [Supported by NIH]*

1. Introduction

In this paper we introduce a sequence of data processing procedures that allow the extraction of various types of data from real-time magnetic resonance (MR) images. The data are intended to be used in speech production studies. Examples of data types of interest are geometric measurements between articulators, such as degree of constriction, as well as articulator positions relative to fixed anatomical landmark, such as the larynx location with respect to the vertebrae, and articulator shapes, such as tongue shape. Also the vocal tract aperture function and area function are oftentimes of interest to the speech production researcher since they can serve as a basis for tube models that allow the approximate computation of the acoustical transfer function and the resonance frequencies of the vocal tract.

In this paper, we present various examples of real-time MR images taken from recordings for our team's (sail.usc.edu/span) speech production studies. With an image reconstruction rate of 22 frames per second, large amounts of image data must be processed, and it is desirable that this processing be automated and requiring as little human interaction as possible, for both efficiency and consistency. Moreover, even fully automatic image processing by a computer can be very time consuming, and we are forced to look for computationally efficient solutions even if those may compromise the achieved accuracy to some small extent. Requirements for the algorithms are robustness towards noise and MR motion artifacts, as well as robustness towards (or correction of) head movement of the subject.

The following sections of this paper describe considerations for MR imaging and reconstruction, the procedures for teeth contour pseudo-implantation, articulator contour tracking, and midline and aperture function calculation. The final section uses a specific example to illustrate and validate the proposed methods using a comparison of the vocal tract resonance frequencies derived from an MR image with those obtained from a simultaneous synchronized noise-cancelled audio recording.

2. Data Collection and MR Image Reconstruction Considerations

In general, the MR image acquisition is constrained by a direct tradeoff between temporal versus spatial accuracy. While static postures of the vocal tract can be captured with high spatial accuracy even in 3D using multi-slice cine MRI techniques (Story et. al. 1998), a true real-time analysis suffers from the relatively slow speed of today's MR technology when compared to x-ray or ultrasound imaging. We therefore chose to consider only fast single-slice midsaggital MR images.

While the particular temporal and spatial resolution parameters are chosen depending on the goals of the study, the images shown in this article were acquired with a repetition time of TR = 6.5ms on a GE Signa 1.5T scanner with a 13 interleaf spiral gradient echo pulse sequence (Narayanan et. al., 2004). The slice thickness was approximately 3mm. A sliding window reconstruction at a rate of 22 frames per second was employed. The field of view (FOV) was adjusted depending on the subject's head size and other constraints outlined below. The example images shown here cover an area of 18.4cm by 18.4cm at a resolution of 68 by 68 pixels. It should also be noted that the subject is always in supine position during the scan, but the resulting images have been rotated into upright position.

It is certainly clear that for a more exact temporal analysis a higher frame rated is desired. However, the higher frame rate is achieved at the expense of a lower spatial resolution, and hence the images would tend to look blockier with the results that any derived geometric measurements are less exact. Hence, the overall achieved accuracy is a compromise between adequate temporal and spatial resolution. The MR operator can try to make improvements by zooming in on the vocal tract more, i.e. decreasing the FOV, but at some point that will introduce MR typical spatial aliasing, and it also means the loss from the image of robust anatomical landmarks such as the vertebrae that maybe needed for head movement correction. Hence, the MR parameter selection is not trivial and requires some experience and experimentation.

3. Teeth Contour Pseudo-Implantation and Head-Movement Correction

The MR process only images the hydrogen concentration within the object inside the scanner. Since both bones and teeth have very low hydrogen content, they show up black in the image, as does air; whereas tissue is gray or white.

However, for some experiments such as the study of fricative sounds, the front teeth may be of particular interest. Previous studies have addressed this problem in a number of ways including by superimposing the MR images with an image of the teeth obtained separately using electron beam computed tomography (Story et. al., 1996). We propose a different and probably more convenient method to address this problem. In addition to

the MR recordings for the actual experimental investigation, we also acquire a sequence of images in which the subject presses their lips and the tongue on both sides of the upper and lower front teeth. In those images, the exact teeth outline shows up as a black area since it is completely surrounded by the tissue of tongue and lips. An example of such an image is shown in Figure 1, where the teeth outline has been traced manually with a white line. Furthermore, two anatomical landmarks related to the teeth are selected as anchors for the teeth contours. They are indicated by the black rectangles in Figure 1. For the lower teeth, this area includes the jaw bone; for the upper teeth, it includes some structures of the nasal cavity.

We then carry out a cross-correlation based search for the anchor areas in each image of the actual (main) image recording of the experiment, thereby allowing both translational and rotational displacement. Upon finding the maximum cross-correlation of the anchor areas the teeth contours are now superimposed in the main image sequence. Figure 2 shows an example image with pseudo-implanted teeth contours.

An equivalent process is used to track the location of the spine in order to detect and correct any lateral movement of the subject in the scanner. Figure 1 shows an additional black rectangle around the vertebrae and an additional white line cutting across the vocal tract near the epiglottis. This line will be used to provide an origin for the midline-based coordinate system described in Section 5.

Lastly, it should be noted that the teeth contour pseudo-implantation is an optional process that may not be required for some studies, such as tongue contour analyses. The process takes a few seconds per frames on a PC, and our software tools allow the correction of individual frames by hand.

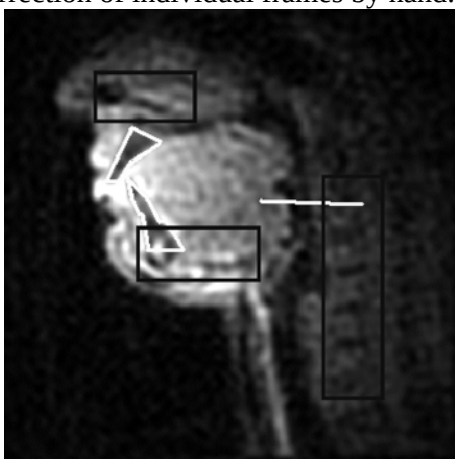

Figure 1. Teeth recording: The white polygons are the manually traced teeth contours, the black rectangles are the associated tracking areas.

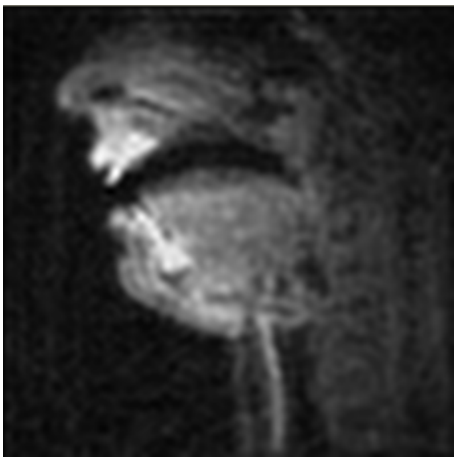

Figure 2. Sample image of the speech sequence with implanted teeth contours.

4. Contour Tracking

The most difficult processing step is the automatic or semi-automatic tracking of contours of interest in the 2D midsaggittal image sequence of the vocal tract. An

example showing the manually traced outline of the complete vocal tract is shown as white dotted lines in Figure 3. It should be noted that some studies may not require information on the entire vocal tract but only certain parts of it such as the tongue for shape analysis. In any case, the manual tracing process is very time consuming and may also introduce human errors. We therefore employ the SNAKE algorithm (Kass et. al., 1987; Lucas and Kanade, 1981; Chan and Vese, 2001) for automatic contour tracking.

Upon manual initialization of the contours of interest in the first frame of the image sequence, the SNAKE algorithm uses an optimization procedure to automatically follow the contours throughout the rest of the sequence. The algorithm uses information derived from the intensity gradient of the images and from optical flow. The latter is a method to estimate frame-by-frame displacement of segments of the image through a correlation-like analysis. The optimization procedure is further constrained by the curvature and continuity of the resulting contour so that the algorithm obtains smooth tracking results.

It should be noted that the standard SNAKE algorithm uses manually chosen weighting factors to blend together the information from the gradient, the optical flow, the curvature, and the continuity constraints during the optimization procedure. While the selection of the weighting factors has to be done only once for each image sequence, we have attempted to implement an automatic weighting procedure, which provides satisfactory results but is still the subject of ongoing research.

A limitation of the SNAKE-based tracking can be seen in images following occlusions between two contours of interest. An example is shown in Figure 4, where some points corresponding to the tongue tip contour are ‘stuck’ to the palate near the upper incisors, and some points of the pharyngeal wall contour are ‘stuck’ to the back of the tongue. The reason for the erroneous tracking results is that in images with occlusions, the intensity gradient in the occluding areas is zero since there are no separating air-tissue boundaries at that moment. When the occluding articulators separate again some tracking points may remain ‘attached’ to the wrong air-tissue boundary. But even though the SNAKE-based tracking results require some manual inspection and correction, the overall processing time is still much improved over a purely manual approach.

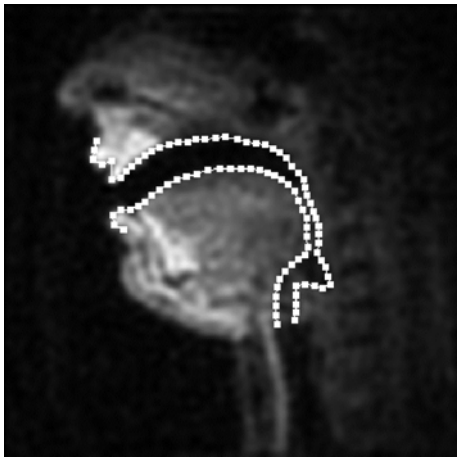

Figure 3. Example of properly initialized and tracked image showing the entire vocal tract outline in white dotted lines.

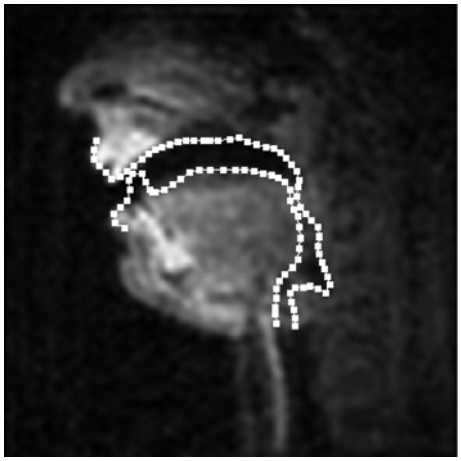

Figure 4. Example of tracking errors after occlusions of tongue/palate and tongue/pharynx.

5. Vocal Tract Midline and Aperture Function Computation

The processing steps after contour tracking may differ depending on the experiment. We want to focus in this section on the special case where the aperture function of the vocal tract is of interest. The aperture function describes the open width of the vocal tract as measured perpendicular to the vocal tract midline from the larynx toward the lips. Conversely, the midline of the vocal tract can be defined as the connecting line of the center points of the vocal tract aperture segments. In the following, we outline a fully automatic procedure to compute the aperture function starting with the construction of the vocal tract midline.

Given the vocal tract outline, two tangent line segments on the contours are found that correspond to the larynx and opening at the front of the mouth. The mid points of these segments serve as the end points of the vocal tract midline. We then recursively carry out the bisection algorithm to find more support points for the midline. In Figure 5, the resulting nine midline support points obtained through three recursive bisections are shown as white x-markers.

As a next step we fit a cubic smoothing spline through the nine support points. It should be noted that, generally, the smooth spline will deviate from the original support points depending on the smoothing factor. We then use the smooth midline spline curve to find finely spaced perpendiculars and compute their intersections with the two vocal tract outline contours. The resulting segments describe the aperture of the vocal tract along the midline. This procedure, however, generally leads to inaccurate results at the mouth opening because of the bend of the smooth midline spline curve. The spline curve does not intersect the mouth opening tangent at a 90 degree angle. We therefore propose as a simple trick the mirroring of the original midline support points on the mouth opening tangent prior to spline fitting. This is shown in Figure 6. Due to symmetry, the resulting cubic smoothing spline (dotted curve) now intersects the mouth opening at a 90 degree angle, and the aperture segments can be found more accurately. The same mirroring

process can simultaneously be applied to the larynx opening thereby destroying the symmetry condition. However, in practice, the mirroring on both ends works fine due to their spatial separation.

Figure 5. Example image showing the vocal tract outline and the midline points after repeated bisection.

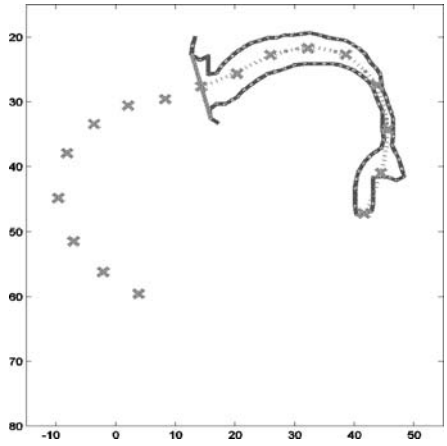

Figure 6. Midline points (x-markers) after mirroring at the mouth opening: guarantees that the smoothed midline (dotted line) is still perpendicular at mouth due to symmetry.

Figure 7 shows the vocal tract outline contours and the aperture segments, which were obtained using the procedure described above. As can be seen, the spacing between the aperture segments is much finer at the mouth and larynx opening in order to more accurately capture the length of the vocal tract. Figure 7 also shows a line connecting the midpoints of all aperture segments. This line is used as the final midline.

While the aperture function can now be measured along the midline independently of the subject's in-plane head movement, some studies may require the exact measurement of the position of articulators within some fixed coordinate system. Examples would be the position of the larynx (larynx height) or the protrusion of the lips. We address this problem by introducing a midline-based coordinate system. Here the midline forms the main axis of the system, and we define an origin on it at an arbitrary position such as slightly above the epiglottis. Now the lips would be at a positive coordinate above the origin whereas the larynx at a negative coordinate below. In Figure 7 the additional, almost horizontal line cutting across the vocal tract at the back of the tongue defines the zero coordinate. We will refer to this line as the origin anchor line. This line is manually chosen and fixed relative to the spine anchoring area mention in Section 3. The final aperture function in the new coordinate system is shown in Figure 8.

Lastly, it should be noted that the midline-based coordinate is expected to be somewhat robust against in-plane head rotation if the origin anchor line is chosen to be close to the larynx. This is due to the fact head movement like nodding does not shift the origin-defining intersection point along the midline but mainly along origin anchor line. Out-of-plane head movement cannot be corrected in our experimental setup, but a secure padding of the subject's head with MRI compatible foam blocks immobilizes the head fairly well.

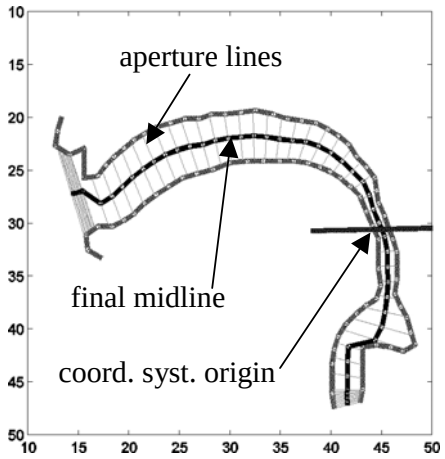

Figure 7. Vocal tract apertures, final midline, origin anchor line.

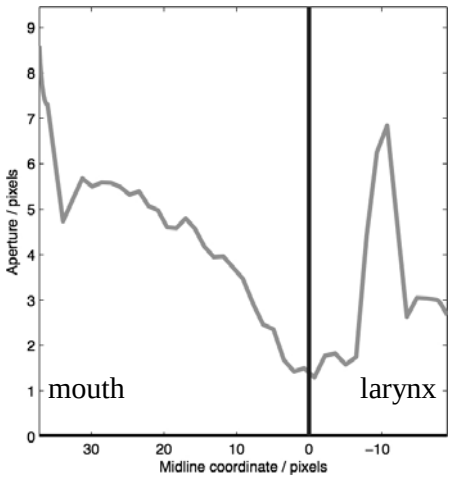

Figure 8. Aperture function: The left end corresponds to the mouth, the right end to the larynx.

6. Example Application: Tube Model and Resonance Frequency Computation

In this section we want to present a brief illustration of our proposed image processing methodology. For this purpose, a single image was extracted from a real-time MR recording of a female subject producing the utterance “la.” The image was taken at a point two thirds into the audio waveform of the syllable, and it contains the vocal tract configuration for the vowel /a/. We then applied the processing steps outlined in the previous sections and obtained the aperture function of the vocal tract.

As a next step, a conversion from the aperture function to the area function was carried out using Ladefoged’s look-up table method (Ladefoged, p.c., 1988). Hereby a scaling step was included in order to achieve a representation in mm instead of pixels. The area function was then fed into the VTAR software (Zhang and Espy-Wilson, 2004) which computed the vocal tract resonances using a lossy tube model. The first and second vocal tract resonance frequencies were found at F1=745Hz and F2=1569Hz.

An independent LPC formant analysis was performed on the synchronized and noise-cancelled audio recording (Bresch et. al., 2006) made during the scan. The formant frequencies obtained from the audio track were F1=959Hz and F2=1411Hz, matching up relatively well with the resonance frequencies derived from the image.

7. Summary

In this paper, we presented processing steps that allow semi-automatic extraction of various types of data from real-time MR image sequences for speech production studies. The step requiring most human supervision and interaction is the contour tracking using the SNAKE algorithm, where occlusions are the cause for most tracking errors.

We also presented a simple method to superimpose the outline of the front teeth into the MR images. This method is conveniently based on a separate MR recording. A related technique can be used for head-movement correction.

We finally outlined an automatic procedure to compute the vocal tract aperture function from the traced vocal tract outline. Such data can be useful for investigating articulatory-acoustic relations.

In summary, the methods described in this paper attempt to provide automatic or semi-automatic solutions to the basic image processing needs for speech production studies that utilize real-time MR image data. Our ongoing work focuses on utilizing the time course information derived from these data processing in studying the dynamics of vocal tract shaping.

References

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. Synchronized and noise-robust audio recordings during realtime MRI scans. To appear in *Journal of the Acoustical Society of America*, 2006.

Chan, T. and Vese, L. Active contours without edges. In *IEEE Trans. Image Processing*, Vol. 10 (2), 2001.

Kass, M., Witkin, A., and Terzopoulos, D. Snakes: Active contour models. *International Journal of Computer Vision*, p. 321-331, 1987.

Lucas, B. and Kanade, T.. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, 115 (4), 2004.

Story, B., Titze, I., and Hoffman, E. Vocal tract area functions for an adult female speaker based on volumetric imaging. *Journal of the Acoustical Society of America*, 104 (1), 1998.

Story, B., Titze, I., and Hoffman, E. Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America*, 100 (1), 1996.

Zhang, Z. and Espy-Wilson, C. A vocal-tract model of American English /l/. *The Journal of the Acoustical Society of America*, Vol. 115 (3), 2004.