Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences

María S. Avila-García, John N. Carter, and Robert I. Damper

***Abstract*—**An important problem in speech research is the automatic extraction of information about the shape and dimensions of the vocal tract during real-time speech production. We have previously developed Southampton dynamic magnetic resonance imaging (SDMRI) as an approach to the solution of this problem. However, the SDMRI images are very noisy so that shape extraction is a major challenge. In this paper, we address the problem of tongue shape extraction, which poses difficulties because this is a highly deforming non-parametric shape. We show that combining active shape models with the dynamic Hough transform allows the tongue shape to be reliably tracked in the image sequence.

***Keywords*—**Vocal tract imaging, speech production, active shape models, dynamic Hough transform, object tracking.

I. Introduction

An important challenge in speech research is to acquire information about the dynamics of vocal tract articulators during real-time speech production. This is very difficult because in general the vocal tract is not easily accessible, the biological structures involved are complex, and the articulators move relatively fast.

Magnetic resonance imaging (MRI) is one modality (originally devised for medical imaging) that offers several advantages and has been widely used by speech researchers [1],[2]. It is based on the application of magnetic fields and radio frequency (RF) pulses to image a desired slice of the human body. Some advantages of this method are the good contrast of soft tissue of which most vocal tract articulators are formed, the possibility of scanning in any plane and the absence of any known hazards to the subject from the magnetic fields used in MRI.

Magnetic resonance images the hydrogen distributed within the body. As such, structures like bone, teeth and air can not be seen. However, the major disadvantage of MRI for speech research is the long scanning time (1-2 seconds), which makes it difficult to obtain information about the movement of the articulators during speech, especially for obstruent sounds (like stop consonants) where the articulators move relatively fast. The tongue is the articulator that moves and deforms the most. As such, it is of much interest to speech researchers. Because of this extensive deformation, however, it can not be easily defined by a parametric shape, and this compounds the problem of visualising this structure in image sequences.

But speech involves very rapid movement of vocal tract structures. When the topic of interest is the dynamic behavior of the articulators, dynamic MRI can be used. These studies attempt to acquire image data while the subject is speaking in a natural way. Recent advances in MRI technology have led to the development of real-time imaging with rates of between 5 [3] and 9 [4] images per second. However, these need expensive machines as well as requiring very sophisticated imaging techniques.

The Southampton dynamic magnetic resonance imaging method (SDMRI) reported in [5] is an alternative technique which achieves an *apparent* high temporal resolution suitable for dynamic studies, by averaging and reordering images acquired over many repetitions of the test tokens. In this paper, we will first describe SDMRI, and then detail how the reconstruction has recently been improved relative to the original version of the method. Thereafter we will introduce a new method for extracting and tracking highly deformable objects such as the tongue.

II. Southampton Dynamic Magnetic Resonance Imaging (SDMRI)

SDMRI consists of acquiring, simultaneously, MR images and the speech data of a subject, typically speaking a nonsense word chosen for its phonetic interest. It works by post-processing the collected images and the corresponding audio data—see [5] for full details. Raw images are generated by the MRI scanner in *k*-space, i.e., the Fourier space. Audio data and raw *k*-space images are manually synchronised, defining a phase for each row of images. Then rows with the same phase generate a new image. Once the images are synchronised, an inverse Fourier transform is applied to generate the final image. SDMRI was implemented to reconstruct multiplanar images of the vocal tract with an apparent sampling rate of 63 Hz, increasing the actual sampling rate of the scanner used in the experiments by a factor of 136.

Because the synchronisation of speech and audio data is manual and so only approximate, and because the different repetitions of the nonsense word are never exactly identical, some parts of the speech signal have no corresponding image data whereas others are over-represented by multiple rows. So SDMRI must solve the problem of missing and multiple rows in the synchronised frames.

Manuscript received 10 November 2004. María Avila-García is supported by a scholarship from the Mexican National Council for Science and Technology (CONACyT).

María S. Avila-García, John N. Carter, and Robert I. Damper are with the School of Electronics and Computer Science, University of Southampton, Hants. SO53 2AB, UK (corresponding author J. N. Carter, phone +44 023 80 592405; fax +44 023 80 594498; email jnc@ecs.soton.ac.uk).

The original algorithm defined the phase by linear stretching or dilation of the total duration of each utterance (token). This approach inherently assumes that each phone duration is identical, i.e., it represents the same percentage of the total duration of the complete token. The new implementation used here calculates the phase by linear stretching or dilation of the duration of each individual phone It results in better definition of the tongue and lower lip boundaries in most cases, as shown in Figure 1.

When the synchronised images have some missing rows in *k*-space, the final reconstructed images will be blurred and the boundaries and edges in the image are not very well defined. The original algorithm solved this problem by filling the missing rows from a simple average of the neighboring frames. Although more sophisticated algorithms exist [6],[7] and have been investigated by us for this application, the original solution of [5] has not been bettered.

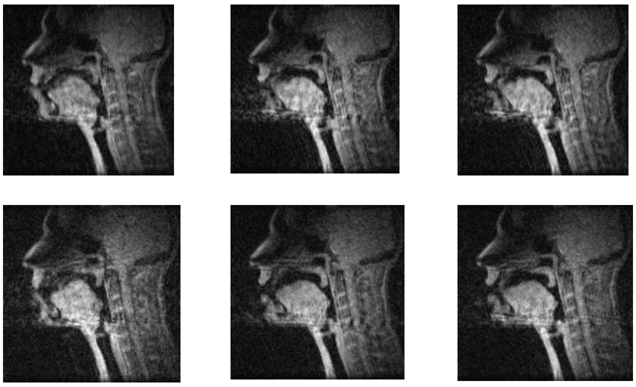

Figure 1. Results of the original synchronisation algorithm (first row), and of the new implementation (second row). There is a general improvement in visual quality (but not in all cases).

III. Shape Extraction

For the automatic extraction of the tongue shape from SDMRI images, three key aspects must be considered.

1. The images are very noisy.
2. The tongue shape is highly deforming and cannot easily be represented by a parametric model.
3. It is advantageous to avoid any need for initialisation of the shape extraction procedure at some particular region of the image, since this compromises fully-automatic implementation.

An anisotropic filter as described in [8] was applied to help solve the first problem This filter gives the advantage of reducing the noise while also preserving the edge information, as shown in Figure 2.

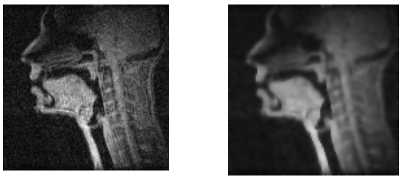

Figure 2. Result of the application of an anisotropic filter (right) to a noisy SDMRI image (left).

There are many different techniques for automatic shape extraction described in the literature. All have their particular advantages and disadvantages.

Snakes and their various extensions [9] have been shown to be very effective in favourable cases (e.g., when the shape to be extracted is smooth), but they usually require some form of initialisation, which is generally difficult to do automatically.

The Hough transform [10] is also highly effective and, being evidence-based, requires no initialisation. The disadvantage is that a good parametric model of the target is required and computational complexity can be a problem for models with many parameters.

Active shape models (ASMs) [11] have proven to be very effective in learning to parameterise deformable shapes. However, in their original implementation, they depend on an optimisation procedure to fit the features. This also may suffer from initialisation problems, especially in complex images.

In previous work, a model of the tongue was manually traced using 39 different tongue shapes extracted from a set of hand-labeled images [12]. This has been used to generate an active shape model as follows. First, each shape had to be delineated to the same anatomical region. To achieve this, the centroid of the shape was calculated first, then a vertical axis was drawn and an angle θ used to define the set of points to be discarded—see Figure 3(a). A value for θ of 45° has given good results. This process was performed for each shape in the training set. Figure 3(b) shows an example of an original tongue shape, and Figure 3(c) shows the reduced tongue shape.

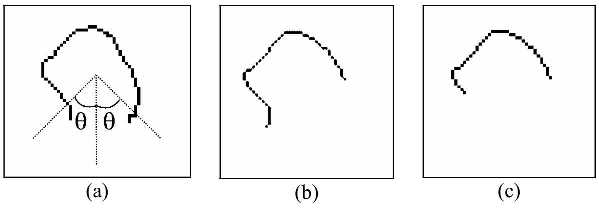

Figure 3. (a) Selection of the tongue shapes, (b) original tongue shape, (c) reduced tongue shape.

Each shape in the training set must be described by the same number of points, defined here as the length in points of the longest training example (61 in this work). Interpolation and smoothing were applied to those shapes that were represented with a smaller number of points. Then the tongue shapes were aligned and a mean shape, $\bar{x}$, was calculated. Alignment was performed by scaling, translating, and rotating the tongue shapes as in [11]. Fig. 4(a) shows the 39 original tongue

International Science Index, Medical and Health Sciences Vol:1, No:2, 2007 waset.org/Publication/9680

shapes superimposed and Fig. 4(b) shows the 39 aligned tongue shapes.

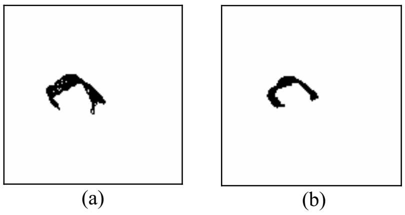

Figure 4. Alignment results. (a) Original tongue shapes superimposed. (b) Aligned tongue shapes superimposed.

The next step is to perform principal component analysis (PCA) and calculate the covariance matrix. The principal axes of the set of aligned shapes are described by the eigenvectors of this matrix, which describe the most significant modes of variation modifying different parts of the tongue. The results of PCA are presented in Table I.

TABLE I
SET OF EIGENVALUES OF THE COVARIANCE MATRIX DERIVED FROM THE ALIGNED TONGUE SHAPES

λ	Eigenvalue	Percentage	Accumulated percentage
λ_1	87.39	72.2	72.2
λ_2	21.85	18.0	90.3
λ_3	2.85	2.3	92.7
λ_4	1.57	1.3	94.0
λ_5	1.35	1.1	95.1
λ_6	1.11	0.9	96.0

The model of the tongue is then defined as:

$$x = \bar{x} + Pb \tag{1}$$

where $\bar{x}$ is the mean shape, P is the matrix of eigenvectors and b is the column vector of the corresponding eigenvalues. In this early work, as proof of concept and to limit computation time, we have considered the first eigenvalue only. Hence, b is a scalar, and P is a one-dimensional vector.

Rather than fitting the shape to the image by a computationally-expensive iterative refinement as in [11], we have formulated a new technique based on the Hough transform (HT), which uses the active shape model in an evidence-gathering mode. The strong advantages of the HT are that it requires no initialisation and avoids iterative search. In its simplest form, it searches for deformable shapes over the whole image, neglecting changes in size and orientation. Its main disadvantage is that the voting space becomes very large for a model with many parameters, and this is the reason for limiting initial work to the first eigenvalue only.

Starting with an edge image, for each edge pixel all possible centroids of all possible shapes are added into a multidimensional accumulator space whose dimensions are x, y, and the eigenvalues. Once all pixels have been considered, the largest peak in the accumulator is chosen as the solution.

We have tested this using the tongue model with (as mentioned above) just one eigenvalue. The dimensionality of the accumulator space is $128 \times 128 \times 11$ for each frame, where 128×128 is the size of the image and 11 is the number of steps into which we discretise the first eigenvalue. In consequence, we fit 11 different tongue shapes. The shape parameter is (as usual with active shape models) allowed to vary in the range $-3\sqrt{\lambda} \leq b \leq 3\sqrt{\lambda}$.

The implementation of the HT considered the tongue shape as a set of points and not as a parametric shape. Because of the sparse data available to us (collecting MRI data is very expensive), the set of testing images is the same as the training set from which the model of the tongue shape was defined. We believe this is justified in this initial work, since if our methods do not work on the training data, they will hardly be worth developing further.

Results are shown in Figure 5. The first row shows three of the original edge images, and the second row shows superimposed the results of applying the new HT algorithm (incorporating an active shape model) to extract tongue shape. Results for the first and second frame are quite good, defining not only a good position of the tongue but also a good approximation of its shape. However, the third result was not especially good: both the position and shape selected by the algorithm are poor, judged visually. We surmise that this is because the images are analysed in isolation.

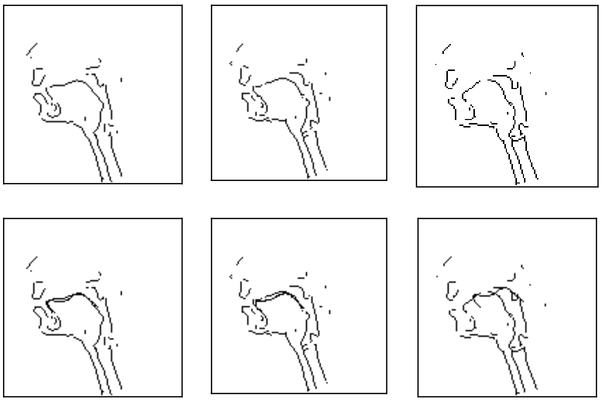

Figure 5. The first row shows the original edge image obtained from SDMRI. The second row shows the results of extracting tongue shape using the new DHT+ASM tongue model, superimposed on the edge image.

To exploit information contained in other frames in the sequence, we propose to combine ASMs with the dynamic Hough transform.

IV. Dynamic Hough Transform (DHT)

The DHT algorithm has been demonstrated in [13], [14] to give excellent results for tracking known objects with arbitrary velocity by finding a smooth trajectory across the whole image sequence using dynamic programming. The implementation of this algorithm (but using now the generated tongue model) involves the inclusion of another constraint: smoothness in the deformation of the tongue.

The original algorithm maximises the energy function:

$$E = w_1E_1 - w_2E_2 - w_3E_3 \quad (2)$$

where E_1 is the sum of the evidence peaks in the accumulator across the sequence, E_2 is the sum of the smoothness terms and E_3 is the sum of the velocity terms—see [13],[14]. We have added a fourth constraint, namely the sum of the deformation of the tongue shape (i.e., the sum of the differences of the eigenvalues for adjacent point in a trajectory). Thus, the algorithm returns the optimal shape in each image.

A first implementation of this algorithm has been done in Matlab. The efficiency limitations of Matlab necessitated using just the eigenvalue with the most significant mode of variation, discretised into only 11 different values, as previously described. The optimisation first applies the new version of the Hough transform to each frame in the sequence individually. Then a second pass finds the optimal sequence of peaks (or 'near peaks') by dynamic programming. Since the tongue only deforms and does not translate, the search was carried out over a small change of velocity of ±2 pixels per frame.

Results are shown in Figure 6. As previously (Fig. 5), the first row shows the original edge image, and the second row shows superimposed the results of using the new DHT+ASM tongue-extraction algorithm. As can be seen, the shape is well defined and matched with the original edge image. Results are clearly superior to those of Figure 5.

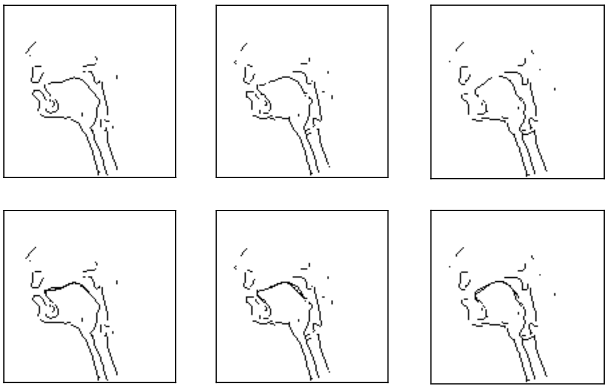

Figure 6. The first row shows the original edge image obtained from SDMRI. The second row shows the tongue shape extracted using the new DHT algorithm with the ASM tongue model, superimposed on the edge image.

V. Conclusions

We have described work aimed at the automatic extraction of information of vocal tract shape and dimensions from magnetic resonance images acquired during speech production. The SDMRI method can be used to infer dynamic information about rapid changes in articulator shape and position, apparently increasing acquisition rate by a factor of 136. A new shape extraction algorithm which combines an active shape model with the dynamic Hough transform has been introduced and described. Experimental results reveal that this new algorithm is very effective in tongue shape extraction from a sequence of images. The key is to use the complete sequence to constrain the search in Hough space so as to find a global optimum. We conclude that combining the established computer vision techniques of active shape models and the dynamic Hough transform does offer advantages in this application and gives promising results.

For the immediate future, the current rather slow Matlab software will be reimplemented in C to decrease the computational time, making the method more practical. As with other applications of the DHT, this algorithm is expected to maintain good performance when some information of the tongue is missing, although this remains to be tested. We also intend to test the algorithm with 'unseen' images, to verify the suitability of applying a model of the tongue found from one subject to a different subject.

References

[1] T. Baer, J. C. Gore, S. Boyce, and P. W. Nye, "Application of MRI to the analysis of speech production", *Magnetic Resonance Imaging*, vol. 5, pp. 1-7, 1987.

[2] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimension using MRI: Vowels", *Journal of the Acoustical Society of America*, vol. 90, No. 1, pp. 799-828, 1991.

[3] D. Demolin, S. Hassid, T. Metens, and A. Soquet, "Real-time MRI and articulatory coordination in speech", *Comptes Rendus Biologies*, vol. 325, No. 4, pp. 547-556, 2002.

[4] S. Narayanan, K. Kayak, S. Lee, A. Seit, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production", *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp.1771-1776, 2004.

[5] M. Mohammad, "Dynamic measurements of speech articulators using magnetic resonance imaging", Ph.D. thesis, Department of Electronics and Computer Science, University of Southampton, Southampton, UK, 1999.

[6] J. Roerdink, and M. Zwaan, "Cardiac magnetic resonance imaging by retrospective gating: Mathematical modeling and reconstruction algorithms", *Journal of Applied Mathematics*, vol. 4, pp. 241-270, 1993.

[7] B. Jennison, and J. Allebach, "Maximum likelihood image reconstruction from Fourier-offset data using the expectation-maximization algorithm", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, vol. 4, pp. 2597-2600, 1991.

[8] P. Perona, and J. Malik, "Scale-space and edge detection using anisotropic diffusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629-639, 1990.

[9] A. Blake, and M. Isard, "Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion", Springer, Berlin, 1998.

[10] J. Princen, H. J. Illingworth, and J. Kittler, "A formal definition of the Hough Transform: Properties and relationships", *Journal of Mathematical Imaging and Vision*, vol. 1, pp. 153-168, 1992.

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models – their training and application", *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-59, 1995.

[12] M. Mohammad, E. Moore, J. N. Carter, C. H. Shadle, and S. R. Gunn, "Using MRI to image the moving vocal tract during speech". *Proceedings of Eurospeech '97*, Rhodes, Greece, vol. 4, pp. 2027-2030, 1997.

[13] P. Lappas, J. N. Carter, and R. I. Damper, "Object tracking via the dynamic velocity Hough transform", *Proceedings of IEEE International Conference on Image Processing*, Thessaloniki, Greece, pp 371-374, 2001.

[14] P. Lappas, J. N. Carter, and R. I. Damper, "Robust evidence-based object tracking", *Pattern Recognition Letters*, vol. 23, no. 2-3, pp. 253-260, 2002.