MULTI-PLANAR DYNAMIC MAGNETIC RESONANCE IMAGING: NEW TOOLS FOR SPEECH RESEARCH.

Christine H. Shadle, Mohammad Mohammad, John N. Carter and Phillip J.B. Jackson.
Department of Electronics and Computer Science,
University of Southampton, Southampton, SO17 1BJ UK.

ABSTRACT

A multiplanar Dynamic Magnetic Resonance Imaging (MRI) technique that extends our earlier work on single-plane Dynamic MRI is described. Scanned images acquired while an utterance is repeated are recombined to form pseudo-time-varying images of the vocal tract using a simultaneously recorded audio signal. There is no technical limit on the utterance length or number of slices that can be so imaged, though the number of repetitions required may be limited by the subject's stamina. An example of [pasi] imaged in three sagittal planes is shown; with a Signa GE 0.5 T MR scanner, 360 tokens were reconstructed to form a sequence of 39 3-slice 16 ms frames. From these, a 3-D volume was generated for each frame, and tract surfaces outlined manually. Parameters derived from these include: palate-tongue distances for [a,s,i]; estimates of tongue volume and of the area function using only the midsagittal, and then all three slices. These demonstrate the accuracy and usefulness of the technique.

1. INTRODUCTION

Magnetic Resonance Imaging (MRI) has been used to good effect in speech research, because of the quality and resolution of its soft-tissue imaging and the possibility of imaging any plane or planes. However, the long time taken to acquire a three-dimensional picture has limited its use to relatively static configurations [1,2,3]. Several different research groups have developed different techniques enabling MRI to be used to image the moving vocal tract. Some of these depend on using fast MR scanners [4] or tagging MRI [5]. Others use synchronizing methods built in to the MR scanner [6,7,8]. Our system uses post-synchronization to build up pseudo-time-varying midsagittal images of the vocal tract during normal speech [9]. In this paper we describe the extension of this early work to a multi-planar dynamic MRI protocol, present results showing three sagittal pseudo-time-varying slices, and describe ways of visualizing the data that result.

2. DYNAMIC MRI

To understand Dynamic MRI it is necessary to discuss how MRI systems work. In essence, by the application of suitable magnetic fields and radio frequency radiation, the MR machine captures a signal relating to the spatial distribution of hydrogen atoms in the subject. Since the distribution varies with the type of tissue, the soft tissues are clearly delineated. MR machines collect images sequentially, row by row, but in the Fourier domain [10] rather than the spatial domain. This raw data is commonly known as a K-space image. The time taken to collect each row is short compared to the overall scan time; the time at which each measurement is made is easily determined.

If images are collected synchronously with an external stimulus, pseudo-moving images can be constructed, provided that the underlying motion is repetitive. Most MR machines have this capability, and it is extensively used for cardiac studies where the external stimulus is derived from a patient's heart beat. This technique has been adapted for vocal tract imaging. In one study [6] the subject synchronized their speech to their own heart beat; Masaki et al. [7,8] utilized a repetitive trigger, fed to both subject and machine. In both studies, the subject was required to synchronize to the MR machine, which then performed the image reconstruction automatically. This method is highly constraining, allowing only very short speech sequences, and making no allowances for irregularity in the subject due to mistakes, breathing or lack of training.

Dynamic MRI overcomes these limitations by performing reconstruction offline, after the recording session. This is made possible by careful analysis of a simultaneous audio recording where key components of the speech sequence are identified, their time relative to each scan determined, and the associated position in Fourier space determined.

In our system the subject is scanned while repeating the test phrase as many times as possible. In each such take a few K-space images will be collected, although the automatically generated images will be blurred because of the articulator movement. By repeating the process, i.e. by obtaining many takes and randomizing the start of the scan relative to the start of the utterance, many such K-space images will be obtained, but with each row of an image corresponding to a different time during the utterance.

A simultaneous audio recording is made, and although this is dominated by the machine noise it is possible to recognize and label the speech sequence. This is used to perform an off-line analysis whereby the K-space data is reordered to reconstruct a series of pseudo-frames, synchronized with the average speech waveform. Finally an inverse Fourier transform is applied to each frame, and a sequence of static images created.

The primary advantages of off-line reconstruction as outlined above are: variation in speaking rate can be accommodated; mistakes made by the subject such as running out of breath during a take can be dealt with; a more natural speaking environment results as the machine is synchronized to the subject. A more detailed description of this process can be found in ref. [11].

The lower limit determining the minimum time resolution, or frame length, possible is T_R, the time taken to acquire one

row of a K-space image. In our previous study, T_R=21ms; in this study T_R =16ms. It would seem that the number of takes should depend simply on the length of the speech utterance relative to the time to acquire one row, and the number of rows per image. This would predict that if N frames of length T_R are needed to show an utterance of length $N*T_R$ then N takes are needed. In practice some rows of particular time-frames are acquired twice, others are missed. Duplicates are averaged, thus improving that row's signal-to-noise ratio; missing rows are borrowed from adjacent frames. It has been shown by simulation and experiment that 1.5*N takes are needed for an acceptable reconstruction quality.

The time resolution obtained with Dynamic depends on the particular scanner used. For example, the SIGNA General Electric scanner used in this work was capable of recording a 128 by 128 pixel scan in 2.2 seconds with an acceptable signal to noise ratio. More modern machines are capable of faster scan rates and will generally also have higher signal-to-noise ratios.

3. EXTENSION TO 3D.

We first reported our dynamic MRI technique in 1997 [9]. In that study only the midsagittal slice was imaged. For each take that slice was scanned repeatedly for 2.8 s while the subject repeated the nonsense word [pasi]. On average 6 tokens were said during each 2.8 s scan. A total of 60 takes were scanned, resulting in 25 reconstructed frames with a resolution of 21 ms.

Extending this to 3-D was relatively easy. Each sagittal scan was replaced by a continuous sequence of left, middle and right sagittal scans, where the left and right imaged planes were parallel to but offset from the plane of the middle scan. The total scan time was nearly three times the scan time for a single slice, and as a consequence the subject was asked to repeat the speech segment for longer.

Although a number of different scan sequences were considered, it was felt that interleaving the acquisition of the three slices was preferable to acquiring all images for one slice contiguously. Thus any long-term articulatory changes due to the subject tiring would influence all three slices equally.

The fundamental limit to the efficacy of this technique depends only on the subject's stamina. Increasing the number of slices increases the time per take; increasing the length of the utterance decreases the number of repetitions per take and thus increases the total number of takes. Either change thus increases the total scan time required of the subject.

4. RECORDING AND IMAGE ACQUISITION

A SIGNA General Electric scanner with a field strength of 0.5 Tesla was used for this study. A volume 240mm by 240mm by 27mm was imaged, in 3 slices with a fast RF-Spoiled Gradient-Echo sequence. The resulting 128 pixel by 128 pixel images corresponded to a slice 5mm thick in the mid-sagittal plane and similar slices displaced 11mm (center to center) to the left and the right. The T_R for this sequence was 16 ms and the three-slice sequence acquisition time was 6.6 s.

The subject for this experiment, PJ, was an adult male, a native speaker of British English, with normal speech and some phonetic training. The utterance chosen was [pasi], because it is short and requires extensive articulator motion. Using the built-in intercom the subject was prompted to begin when ready. He began repeating [pasi] and the machine was turned on at a random time after the first token. The subject kept repeating [pasi] until he heard the machine stop. On average, 12 tokens were uttered during each scan. The K-space data were then saved before the next take. Twenty-four scans were collected for a total of about 24*12 repetitions of [pasi]. For comparison, scans with exactly the same set-up were collected for [a,s,i], each sustained for 6 seconds.

Sound was recorded on a Sony Walkman Pro analog cassette recorder by placing its microphone next to the intercom speaker built in to the scanner monitor. Sound quality thus obtained is poor, owing both to the intercom characteristics and the high-amplitude scanner noise. However, the recording quality is sufficient for the segmentation necessary for the post-processing, and a microphone inside the scanner (tried earlier) causes image artifacts. [SOUND S0830.WAV]

A separate recording was made of the same subject in a sound-treated booth the day before the MRI session. He was recorded saying the same corpus sitting up and lying down, using a Bruel and Kjaer 4133 microphone and 2636 measurement amplifier to record onto audio cassette tape.

5. VISUALISATION

Post-processing of the acquired K-space images and audio tape recording proceeded as outlined above to generate 39 images, a pseudo-time `movie' with 16 ms framerate, for each of the three slices.

Once the three slices for each pseudo-time frame had been generated, the next step was to outline the vocal tract on each reconstructed image. Automatic feature extraction methods have been tried but have not proven effective. Manual outlining was performed on all 39 frames. One advantage of manual outlining is that anatomical features can be labeled as they are traced.

No special tools are required for this. The outline of the vocal tract was simply drawn onto each image, using different colours to indicate different parts of the anatomy. Specific conventions are adopted to label possible ambiguous areas, such as where the tongue is in contact with the palate or where ghost outlines indicate that an anatomical structure does not completely occupy the slice. This coding scheme allows different segments of the anatomy to be selected and linked together automatically to form meaningful outlines. For example, a segment where the tongue is in contact with the palate is used twice, once in describing the outline of the tongue, and in describing the shape of the palate. The labeled pictures are then automatically processed, and the extracted edge segments used for further analysis. [IMAGE S0830.GIF]

From these labeled data three-dimensional surfaces are reconstructed, by carefully combining adjacent slices. This process is semi-automatic and enables the construction of full 3-D models of the imaged part of the vocal tract for each frame.

Many computer graphics tools exist to draw the 3-D time-varying outlines, once they are constructed. However, very few of these representations are amenable to the human viewer. Rendering surfaces opaque makes it difficult to see into the imaged volume; leaving surfaces in a wire-frame form is confusing.

To make the data useable we have adopted two techniques. First, we build a fully animated vocal tract from our data; second, we process our data to extract information that can be presented in a traditional manner.

The three-dimensional model is constructed automatically. First the parts of the anatomy to be viewed and the method of viewing, e.g. solid or wire frame, are selected. Then the model is constructed using the Virtual Reality Modeling Language. The model can then be viewed with standard WWW browsers equipped with a suitable plug-in module. The model is fully interactive and allows the viewer full control of the viewpoint by controlling the orientation of the model in 3-D or selecting from pre-selected viewpoints. [3D IMAGE S0830.WRL]

The VRML model allows: full control of the model orientation in 3-D; zooming in and out; hiding and revealing anatomical features; and stepping backwards and forwards in time. It allows the spatial and temporal relationships between different parts of the anatomy to be visualized, but is not suitable for making quantitative measurements.

When quantitative measurements are required it is preferable to use the 3-D data to mimic experimental measurements made by other techniques. For example it is possible to synthesize measurements of the position of the velum, jaw and tongue, by modeling the physics of these measurements within the four dimensions of Dynamic MRI.

6. VISUALISING THE TONGUE.

Following our earlier work on Enhanced EPG [12] we have developed an augmented display, which not only shows where the tongue is in contact with the palate, but also gives the vertical separation between the underside of the palate and the top surface of the tongue when not in contact. Figure 1 shows this representation for static and dynamic configurations for [a], [s] and [i], where the darkest regions represent areas of contact, and lightest region represents the maximum distance away. The scale on the right hand side shows the distance from the front of the lips, while the scale along the bottom indicates the distance between upper and lower surface in terms of grey levels. White spaces near the top of each image indicate undefined vertical distance since the teeth do not appear in MRI images.

Although distances were computed in five-pixel chunks along each slice, the information proved much harder to absorb when presented in corresponding blocks of uniform grey level within each block. Low-pass-filtering the blocks proved essential to the perception of a smoothly varying surface.

As expected, tongue-palate vertical distance is largest for [a]. All three sounds show evidence of a tongue groove, that is, the midsagittal slice has a greater tongue-palate vertical distance. For [i] the groove is pronounced only posteriorly, consistent with an arched anterior tongue shape and posterior groove; for [a] there is evidence of an anterior 'dimple' (as described in ref. [13]); for [s], the tongue tip is closest to the

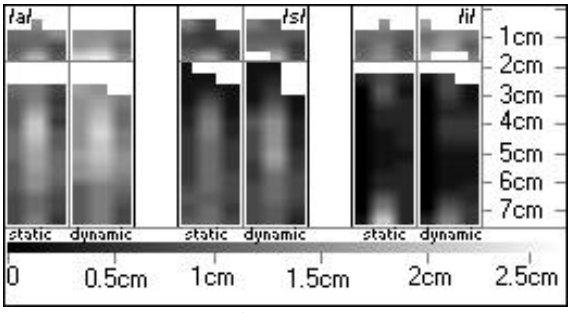

Figure 1 Visualization of the vertical distance between the surface formed by palate and lip, and that formed by the tongue and lower lip. Left, middle and right image pairs show [a,s,i], sustained for the left, extracted from [pasi] for the right of each pair. See text for meaning of scales.

palate, with groove visible for ≥ 3 cm from lips, consistent with an anterior groove. All of these cases are consistent with Stone and Lundberg's four-way tongue-shape classification [13],

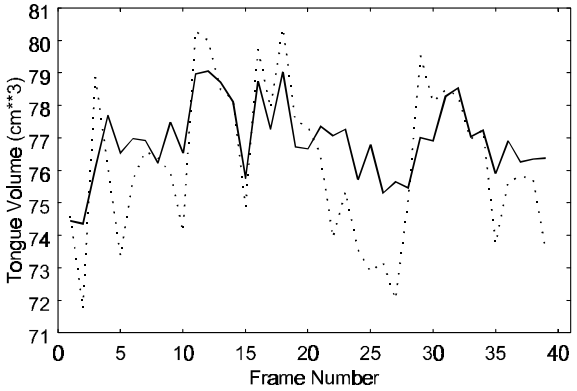

Figure 2 Two estimates of tongue volume for each frame of [pasi], using: midsagittal slice only (dotted line), three sagittal slices (solid line).

although it should be noted that tongue surface shape will correspond but is not identical to vertical distance between the palate and tongue surface.

Anatomical areas and volumes are simple to calculate from the 3-D data. In each plane the labeled data are used to construct an outline of the air space or articulator. From this the volume of interest can be estimated. If only one slice is available, then the volume is calculated as the area times the estimated width of the tract. When multiple slices are available, simplified numerical integration techniques can be applied to calculate the volume, taking the position and width of each slice into account.

It has long been assumed that total tongue volume in a given subject is constant; Stevens gives values averaged over adult males of 110 cm^3, and adult females, 90 cm^3 [14]. To verify the efficacy of our techniques of tract outlining and volume computation, we have estimated the tongue volume for each frame in [pasi] by two methods: $\hat{V}_1$ uses only the midsagittal slice, and $\hat{V}_3$ uses all three slices. For both estimates we used 27 mm, the width of the imaged volume, as our estimated tongue

width, which excluded any tongue that might extend laterally beyond this. We thus predict both our estimates to be a lower bound of the actual tongue volume, and that the estimate should be more constant over time (i.e. with frame number) as we use more slices to form an estimate. As can be seen in Fig. 2, both of these predictions are borne out. $\hat{V}_1$ varies more with frame number than $\hat{V}_3$; their averages and standard deviations across all frames are $\bar{V}_1$ = 76±2.3 cm^3 and $\bar{V}_3$ = 77±1.2 cm^3, both less than the adult male average of 110 cm^3, but of the same order of magnitude. This gives us confidence in our technique.

Motion and changes of shape for other articulators, e.g. velum, chin or lips, can also be measured using similar methods.

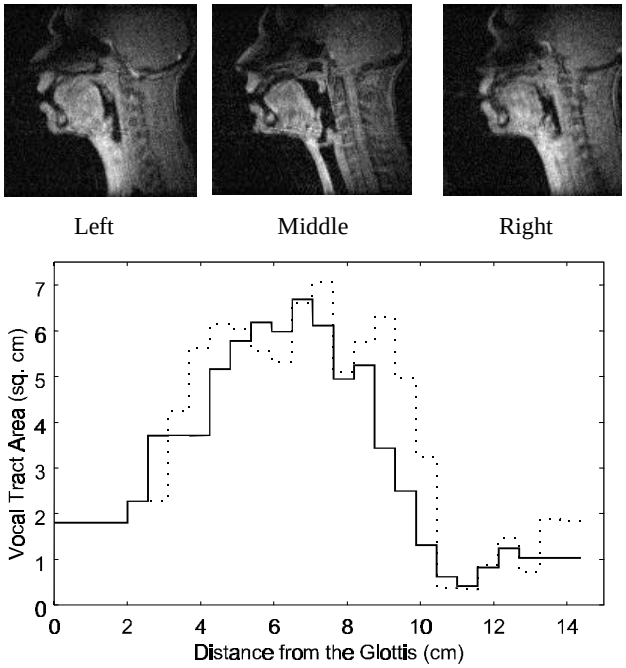

Figure 3 Top: three dynamic images of [i] from [pasi] (frame 31); left, mid and right sagittal scans. Bottom: two estimated area functions for this [i] frame, using: midsagittal distances only and ref.[15] algorithm (dashed line); all three sagittal distances (solid line).

7. AREA FUNCTIONS

Figure 3 shows images for one frame, corresponding to the central 16 ms of [i], and two estimates of the area function derived from these images. The first estimate, $\hat{A}_1$, uses only the midsagittal slice. Midsagittal distances $(d(x))$ are computed along the tract, and $\hat{A}_1(x) = a(d(x))^b$, according to the method described in reference [15]. The second estimate, $\hat{A}_3(x)$, uses all three slices; at 0.5-cm intervals along the tract, a polygon was constructed that intersected the tract-air boundaries in each slice, and its area was computed. This estimate neglects any lateral extension of the tract beyond the 27 mm-wide imaged volume, as well as using a straight-line approximation to the cross-sectional shape within the imaged volume. It is likely therefore that $\hat{A}_3(x)$ somewhat underestimates the true area.

In Fig. 3 the two area estimates differ, though not substantially. Transfer functions were computed based on both estimates and compared to the speech recorded in the sound-treated room. First formant frequencies (F1) matched in all three cases; F2 and F3 were more closely matched by the prediction based on $\hat{A}_3$ (all three slices) than by that based on $\hat{A}_1$ (midsagittal slice plus the Beautemps et al. [15] algorithm).

8. CONCLUSIONS

Dynamic MRI is a potentially useful tool, overcoming the main disadvantages of MRI , the slow image acquisition time leading to static images, while retaining the control of imaging volume. An experiment imaging repeated tokens of [pasi] using three sagittal slices was described. The multiple frames of even this simple data set required development of new methods for viewing the data, which are generalizable to dynamic MRI data of any number of slices and a longer utterance.

REFERENCES

[1] Baer, T., Gore, J.C., Gracco, L.C. and Nye, P.W. (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, *J. Acoust. Soc. Am*. 90, 799-828.

[2] Narayanen, S.S., Alwan, A.A. and Haker, K. (1995) An articulatory study of fricative consonants using magnetic resonance imaging, *J. Acoust. Soc. Am*. 98, 1325-1347.

[3] Shadle, C.H., Tiede, M., Masaki, S., Shimada, Y. and Fujimoto, I. (1996) An MRI study of the effects of vowel context on fricatives, *Proc. Inst. of Acoust*. 18:9, 187-194.

[4] Demolin, D., George, M. Lecuit, V. Mefins, T., Saguet, A. and Raeymaekers, H. (1997) Coarticulation and articulatory compensations studied by dynamic MRI. *Proc. Eurospeech 97*, Rhodes, 43-46.

[5] Stone, M., Lundberg, A., Davies, E., Gullapalli, R. and NessAiver, M. (1997) Three-dimensional coarticulatory strategies of tongue movement. *Proc. Eurospeech 97*, Rhodes, 31-34.

[6] Foldvik, A.K., Kristiansen, U., Kvaerness, J., Torp, A. and Torp, H. (1995) Three-dimensional ultrasound and magnetic resonance imaging: a new dimension in phonetic research, *Proc. ICPhS 95* 4, 46-49.

[7] Masaki S., Tiede M., Honda, K., Shimada Y., Fujimoto I., Nakamura Y. and Ninomiya N. (1997a) Synchronized MRI sampling method for articulatory movement recording, *Proc. for 1997 Spring Meeting of Acoust. Soc. of Japan*, 325-326, Kyoto, 17-19 March.

[8] Masaki S., Tiede M., K., Honda, K., Shimada Y., Fujimoto I., Nakamura Y. and Ninomiya N. (1997b) MRI observation of dynamic articulatory movements using a synchronized sampling method. *J. Acoust. Soc. Am.* 102:5:2, 3166.

[9] Mohammad, M., Moore, E., Carter, J.N., Shadle, C.H. and Gunn S.J. (1997) Using MRI to image the moving vocal tract during speech. *Proc. Eurospeech 97*, 2027-2030

[10] Wright G.A. (1997) Magnetic Resonance Imaging, IEEE Signal Processing Magazine 14:1, 56-66

[11] Mohammad, M. (1999), Dynamic Measurements of Speech Articulators using MRI, PhD. Thesis, University of Southampton.

[12] Chiu, W.S.C., Shadle, C.H. and Carter, J.N.. (1995) Quantitative measures of the palate using enhanced electropalatography. *European J. Disordered Communication*, 30:149--160.

[13] Stone, M. and Lundberg, A. (1996) Three-dimensional tongue surface shapes of English consonants and vowels. J. Acoust. Soc. Am. 99, 3728-3737.

[14] Stevens, K.N. (1998) Acoustic Phonetics. MIT Press, Cambridge, MA.

[15] Beautemps, D., Badin, P. and Laboissiere, R. (1995) Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. *Speech Communication*, 16, 27-47.