USING MRI TO IMAGE THE MOVING VOCAL TRACT DURING SPEECH

M. Mohammad[1] E. Moore[2] J.N. Carter[1] C.H. Shadle[1] S.J. Gunn[1]
[1]Dept. of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
[2]Dept. of Medical Physics, Southampton General Hospital, Southampton SO16 6YD, UK

ABSTRACT

Magnetic Resonance Imaging (MRI) has been used to measure the shape of the vocal tract during speech in several recent studies. Its safety to the subject, high quality imaging of soft tissue, and the ability to select relatively thin imaging planes at any angle are significant advantages over other imaging methods used for speech research. The most significant disadvantage is the long exposure time. As a result most studies have focused on obtaining high-resolution images of the vocal tract volume for static sounds, such as vowels [1], fricatives [5, 6], nasals, the closed phase of plosives [7] and liquids [3,7]. In this paper we will describe our method of obtaining MR images of a moving vocal tract in which we post-synchronise the MR data using a recorded speech signal and thus reconstruct the images without using the MR machine's built-in processing.

1. INTRODUCTION

The Magnetic Resonance Imaging(MRI) modality, as an indirect method of measuring shape, is becoming an increasingly common tool in speech research due in part to the high quality of the soft tissue images and the ability to select imaging planes at any angle with no apparent risks to the subject. However, MRI has the significant disadvantages of not being able to show teeth and bones and haing a very long scanning time. The long scanning time or the low temporal resolution has, until recently, limited speech studies using MRI to the study of static sounds only[1, 5, 6].

While breakthroughs in MRI technology mean that imaging times continue to decrease, single-plane acquisition times are still of the order of a second. Foldvik [2] developed a way to obtain a `movie' of volume images of the vocal tract during production of the diphthong [aI] by exploiting the processing built in for cardiology imaging. Current work at ATR, Japan [4] is focused on a different method of obtaining fine temporal resolution for single-plane images, but that also uses the built-in processing for cardiac imaging.

These studies have concentrated on synchronising the speaker with the MR machine, and used techniques developed for dynamic studies of the heart to produce image sequences during continuous, but repetitive speech. The subject is asked to repeat an utterance many times and to synchronise these repetitions with an external audio signal. The drawbacks of these methods are that the subject has to be carefully trained and that the investigators have no control of the built-in processing of the images. We have developed a new method for measuring the changing shape of the vocal tract, in which we post-process the MRI data using external timing information. By matching the MRI signal with the speech signal recorded the subject during scanning, we are able to reconstruct images for a high-temporal-resolution dynamic representation of the vocal tract with total control of the processing of the images and with no need for an expert subject.

2. DYNAMIC MEASUREMENT USING MRI

Magnetic Resonance Imaging systems measure the density of hydrogen atoms in the material under investigation. However, rather than measuring the spatial distribution of tissue directly, MRI determines the Fourier Transform of the density[9]. Thus in a digitally sampled system the following relationship holds:

$$M_T(K_x,K_y)=\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} m(n_1,n_2)e^{-j(K_x n_1+K_y n_2)}$$

$M_T(K_x,K_y)$ is the recorded magnetisation, known as K-space, and $m(n_1,n_2)$ is the underlying proton density, while N is the size of each dimension of the sampled matrix. The proton or tissue density is easily calculated by taking the DFT of $M_T(K_x,K_y)$ once the magnetisation has been measured.

In general measurement takes place as follows:

1. Appropriate choices of magnetic fields are made to select a 2-D imaging plane or a slice.
2. This region of interest is then excited with a radio frequency pulse.
3. After the pulse has ended, the net excited magnetic moment relaxes and the decay is measured with a suitable radio receiver.
4. This signal is sampled and corresponds to exactly one row in the K-space matrix. The length of this sampling period is governed by the relaxation time of the system, T_R.
5. Additional rows are selected by varying the magnetic field gradient and repeating steps 1-4. A full picture of K-space is completed once all possible rows have been scanned.

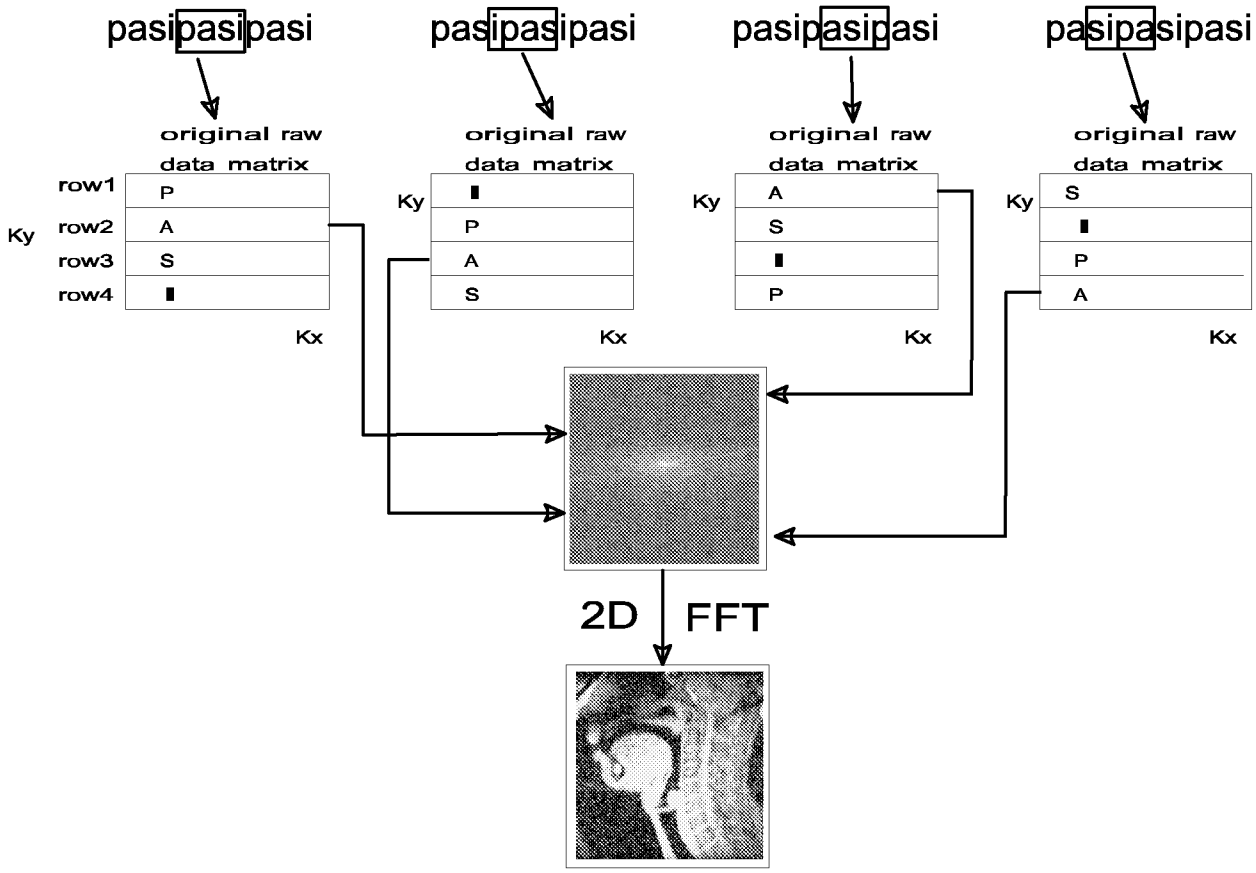

Figure 1 demonstrates in a highly simplified way how rows in K-space are borrowed from different tokens to build up the image for one configuration, [a] in this example.

Normally, for the settings used in this study, steps 1-5 result in a single image of 128 by 128 pixels, which is assembled during 2.8 seconds. The image would be generated automatically from the K-space data which is subsequently discarded. If the subject moves during the 2.8 seconds, a blurred image results. In this study, instead of asking the subject to sustain a single sound for 2.8 seconds, we ask the subject to say multiple tokens of a short ~0.5s utterance, and match the times at which particular rows were imaged to the times at which particular phonemes were uttered.

The reconstructed images thus contain rows generated from different tokens of the same sound. This process is illustrated in Figure 1; the simplifications made for illustrative purposes are that the image has only 4 rows, the utterance has only 4 phonemes of equal duration, and the acquisition time of one row equals the time it takes to produce one phoneme. As shown at the bottom of Figure 1, the synthesised matrix is Fourier Transformed to generate an image.

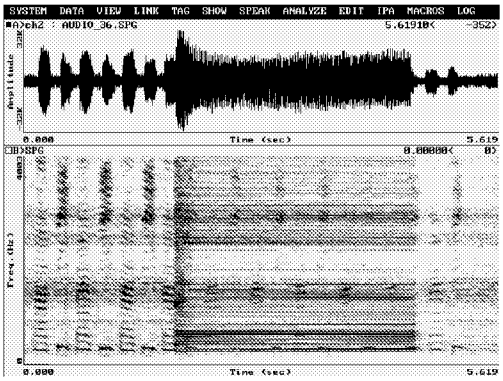

Figure 2 Time domain and spectrogram representation (0-4kHz) of 10 repetitions of /pasi/ during one measurement.

3. METHOD

A SIGNA General Electric scanner with a field strength of 0.5 Tesla was used for this study. A region 200mm by 200mm was imaged with a fast RF-Spoiled Gradient-Echo sequence. The resulting 128 by 128 image corresponded to a slice 5mm thick in the midsagittal plane. The T_R for this sequence was 21 ms and the slice acquisition time was 2.8 s.

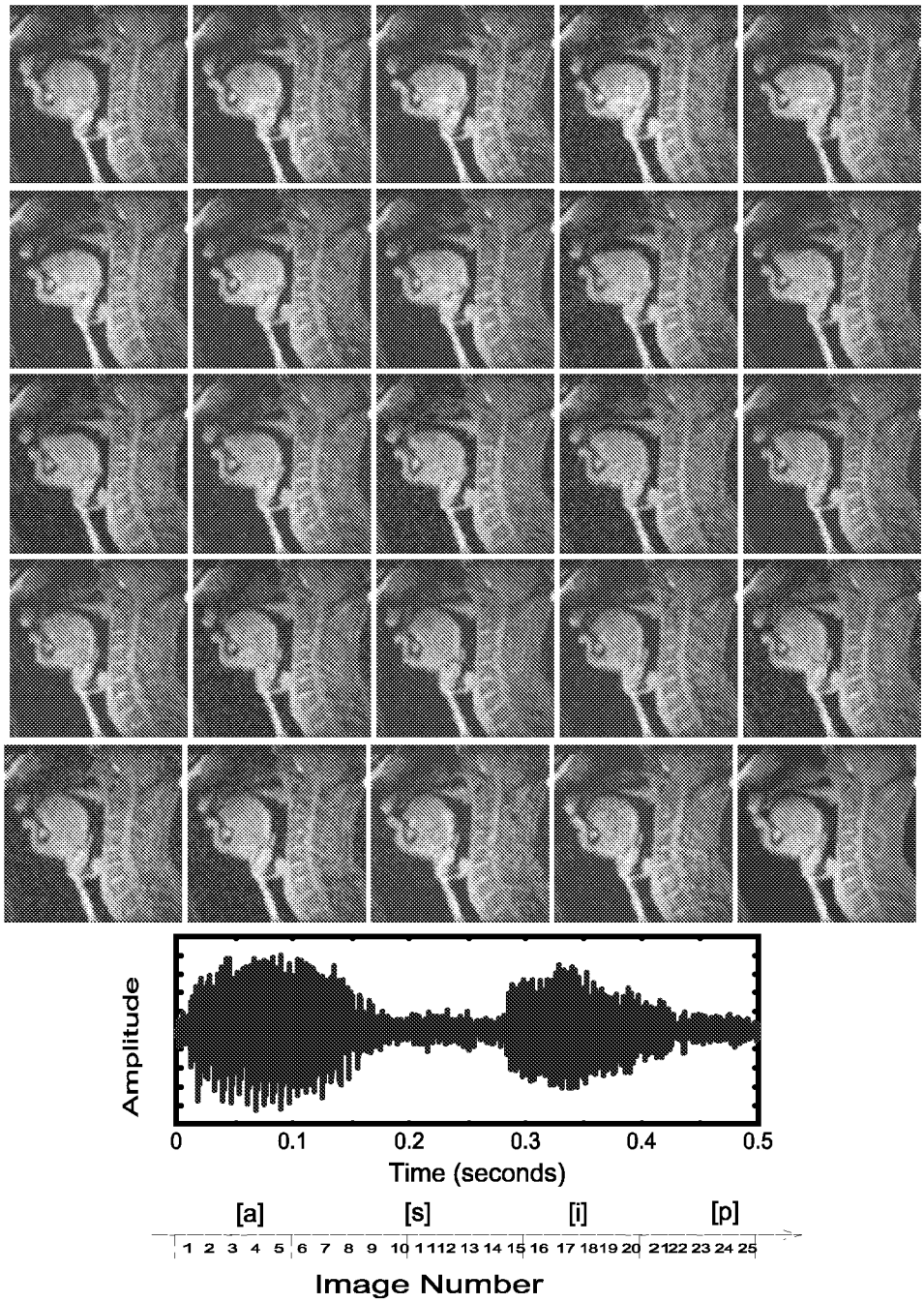

Figure 3 A sequence of MRI pictures for the utterance [pasi], sampled at approximatly 50 Hz. A schematic representation of the audio signal is shown at the bottom. Images are number left to right and top to bottom.

The one subject was an adult male, a native speaker of British English, with normal speech and no 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, 6 tokens were uttered during each scan. The K-space data was then saved before the next take. Sixty scans were collected for a total of about 360 repetitions of [pasi]. For comparison, scans with exactly the same set-up were collected for ? repetitions of each of [a,s,i], while the subject sustained the sound for 2.8 seconds.

The audio signal was recorded by placing the microphone for a Sony Walkman Pro on the monitor station near the loudspeaker of the built-in intercom. The Walkman was left recording throughout the 2 hour long experiment. Placing a microphone inside the magnet results in a higher quality speech signal but was not used because it induced image artefacts.

Figure 2 shows a typical audio signal and its spectrogram. Here, three tokens were said before the machine was turned on. Even with the high-amplitude machine noise, the fornants of the vowels are just visible. By identifying and labelling the beginning and end of each vowel relative to the end of the noise generated by the magnet, the relationship between each component of the audio signal and the rows in K-space were determined.

4. RESULTS AND DISCUSSION

Figure 3 shows a sequence of 25 frames, representing the configurations occurring during a single utterance of [pasi], with an apparent temporal resolution of 21 ms. The movement of the tongue from the low, back position to a high front position is clearly visible. The tongue-tip position further differentiates [s] and [i], and the lips appear closed in the final frame consistent with the [p] of the next token.

Figure 4 shows a comparison of reconstructed images for the sustained [a,s,i] and the frames from [pasi] corresponding to the middle of each of the three phonemes. The spatial resolution is the same in the two sets of images. The most striking difference is in the velum; it is always up for the dynamic case, and is partly down for the sustained [a] and [s]. The pharynx also appears wider in the sustained [i] than in the dynamic [i], and tongue shape differs somewhat for the two [s] images.

Although we must be careful about generalising too much from these difference, they offer a fascinating glimpse of the usefulness of this technique in combining the soft-tissue imaging capabilities of MRI with the naturalness of unstained speech.

In the future we plan to extend the corpus and number of subjects so that we can draw more conclusions about speech production; we also plan to extend this technique to image multiple slices during speech, i.e. 3D imaging.

5. 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] 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.

[3] Masaki, S., Akahane-Yamada, R., Tiede, M., Shimada, Y., and Fujimoto, I. (1996) `An MRI-based analysis of the English /r/ and /l/ articulations,' *Proc. ICSLP-96*, 1581-1584.

[4] Masaki S., Tiede M., K., Shimada Y., Fujimoto I., Nakamura Y. and Ninomiya N. (1997) 'Synchronized MRI sampling method for articulatory movement recording,'
Proceedings for 1997 Spring Meeting of Acoustical Society of Japan, 325-326, Doshisha Univ. Tanabe campus (Kyoto), March 17th-19th.

[5] 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.

[6] 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.

[7] Stone, M., Ong, D. and Lundberg, A. (1996) `An MRI and EPG examintion of /r/ and /l/,' *J. Acoust. Soc. Am.* 100:4:2, 2660.

[8] Story, B.H., Titze, I.R. and Hoffman, E.A. (1996) `Vocal tract area functions from magnetic resonance imaging,' *J. Acoust. Soc. Am.* 100, 537-554.

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

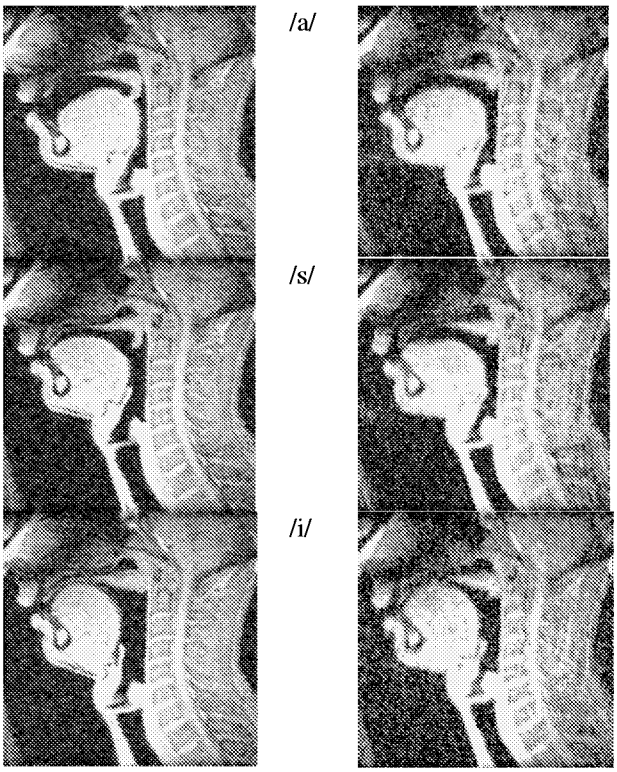

Figure 4 A comparison of sustained(right) and dynamic(left) images for /a/, /s/ and /i/.