*Magnetic Resonance Imaging*, Vol. 5, pp. 1-7, 1987
Printed in the USA. All rights reserved.



573

● Original Contribution

# APPLICATION OF MRI TO THE ANALYSIS OF SPEECH PRODUCTION

T. BAER,* J.C. GORE,† S. BOYCE* and P.W. NYE*

*Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511, USA, †Department of Diagnostic Imaging, Yale School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA

**Computer models of the process of speech articulation require a detailed knowledge of the vocal tract configurations employed in speech and the application of acoustic theory to calculate the sound waveform. Almost all currently available data on vocal tract dimensions come from x-ray films and are severely limited in quantity and coherence due to restrictions on radiation dosage and intersubject differences. We are using MRI techniques to obtain the pharyngeal dimensions of speakers producing sustained vowels. The fact that MRI does not employ ionizing radiation provides speech research with the opportunity to obtain comprehensive bodies of much-needed data on the articulatory characteristics of single subjects.**

*Keywords:* **NMR imaging, Speech, Vocal tract dimensions, Acoustic theory.**

## INTRODUCTION

It is well known that talking involves movement of the speech articulators, including the jaw, tongue, lips, velum and structures in the laryngeal region, and that different configurations of these articulators result in different sounds. According to the *Acoustic Theory of Speech Production*,[2] the vocal tract, or airway between the glottis and the lips, forms an acoustic tube whose shape and dimensions change as the articulators move. Sound is introduced at one end of the tube (i.e. at the glottis for voiced sounds, such as vowels), and it is radiated to a listener from the other end of the tube (the lips). The resonance characteristics of the tube cause it to act as an acoustic filter. As the shape of the tube changes, the characteristics of the filter change, and different speech sounds are produced. Mathematical modelling of this process has been a major focus of speech research, with many potential applications in the fields of speech technology, speech pathology and linguistics. An important component of this modelling effort is the acquisition of basic data on the articulatory configurations encountered in speech production and their concomitant effects on vocal tract shape.

However, due to the relative inaccessibility of the vocal organs, the collection of detailed data on articulator movements and resulting vocal tract shapes is difficult. The necessary instrumentation frequently hinders important elements of the production process, and it permits data to be collected on only a subset of the desired measures and of the types of utterances possible in a language. Typical of this instrumentation is the cine x-ray equipment that was much in use in the early years of speech research. This device can supply data on vocal tract dimensions in the sagittal plane, but cannot provide corresponding data in planes perpendicular to the vocal tract axis. Although, for the latter purpose, the x-ray tomograph has seen some use,[3,6,7] the lateral cine x-ray has been the mainstay of research instrumentation in this field and has provided, by far, the greater bulk of useful data. Now, however, with a greater understanding of the damaging effects of x-rays and with the advent of more stringent regulations associated with their nonmedical use on normal subjects, it is no longer possible to accumulate more than a fraction of the data needed before encountering a subject's dosage limit. Hence today other optical and electronic instrumentation that does not use ionizing radiation is employed in its place, but unfortunately this instrumentation provides the opportunity to observe only a limited

RECEIVED 5/23/86; ACCEPTED 7/25/86.
*Acknowledgements*—This research was supported by NIH Grant Nos. NS-13870 and NS-13617 to Haskins Laboratories and by The Esther A. and Joseph Klingenstein Fund, General Electric Company and the Yale School of Medicine. The authors also wish to thank Dr. Hiroshi Muta for assistance with anatomical identification.

range of articulatory movement. To broaden the experimenter's perspective, several different but complementary data-gathering techniques are usually applied conjointly to the study of an utterance. Then, by piecing together the resulting data, the investigator endeavors to reconstruct all the movements of the participating articulators and their relative timing. Once such a reconstruction has been achieved, its accuracy can be independently verified by applying anatomical knowledge and acoustic theory. The anatomy and the theory can be brought together in a computer model of the articulatory system which is programmed to follow the observed movements and to compute (or synthesize) the resulting vocal tract shapes and acoustic output. The degree to which the synthesized utterance is perceived by a native speaker of the language to be identical to the original utterance is then taken as a measure of the accuracy of the articulatory observations.

We are using a computer model that was originally designed by Mermelstein[9] and subsequently adapted by Rubin, Baer and Mermelstein[11] for use as an interactive tool for the linguistic verification of articulatory movements. Its six key articulators (tongue body, tongue tip, jaw, velum, lips and hyoid bone) are represented in the program together with their mechanical linkages and graphically depicted by the computer as a sagittal cross-section of the vocal tract. The synthesis procedure requires the experimenter to use the physiological observations to specify the positions of the articulators at time intervals corresponding to the period of the glottal cycle (approximately 8–10 ms for male speakers). At each time interval, the computer derives the midsagittal width as a function of distance along the length of the tract, specified every 8.75 mm in the sagittal plane (commencing at the larynx and ending at the lips). Subsequently this function is converted into a corresponding area function or equivalent tubular analog of the vocal tract. The conversion from width to area is achieved by means of a set of functions for different parts of the vocal tract based partly on hypothesis and partly on the best available anatomical data.[5,8,12] Finally, voiced speech is produced by computing the effects of exciting the tube with a waveform whose shape resembles the volume velocity of airflow through the vibrating glottis.

A considerable limitation on the power of the speech synthesis technique as a way of verifying production data lies in the accuracy of the sagittal width to area transformation. At present, most of the available data on vocal tract shape come from a very small number of x-ray studies performed on a limited repertoire of utterances. Moreover, the data relate to different individuals and involve at least four different languages: Japanese,[1] Russian,[2] Swedish[6] and English.[10] Consequently, the data contain many individual differences and inconsistencies. Other data obtained from anatomical studies of morbid tissue are available, but they are usually considered to be a doubtful indicator of the live tract configurations employed in speech. Data have also been collected in vivo using fiberoptic viewing[4] and by making casts.[8] But, because of the obvious experimental difficulties, these data also have limited validity. To gain further precision from the synthesis technique, there is an urgent need to augment and refine the width-to-area conversion data. Therefore, we have turned to magnetic resonance imaging (MRI) as a means of obtaining new data on the relationship between cross-sectional area and length for the vocal tract configurations actively used in speech. To allow adequate time for MRI data collection, we have begun our study by focussing primarily on steady-state (sustained) productions of the point vowels (i, u, a and ae) but have also included some of the intermediate vowels. Although the data we obtain are not yet as precise as those we might in principle gather from x-ray tomograms, the MRI technique is uniquely attractive in its ability to provide an extensive body of dimensional data on the vocal tract of a *single* live speaker without the risks associated with x-radiation.

## METHOD

The subject lies in the supine position on the horizontal patient couch with his head placed inside the rf coils of the MRI unit. To insure that head movements are minimized and that the head can be restored to the same position for each experimental session, a custom-made mold of the head is used. The head mold is mounted on a rigid base that also supports the rf coil assembly. Further stability is gained by a locating wand that makes tactile contact with the subject's nose and induces sufficient proprioceptive feedback to enable him to minimize any residual head motion that might be permitted by the mold. The base itself is attached to the patient couch which, by translation along its major axis, permits a 9.5-cm length of the pharynx to be studied. At one end of this range, the image plane intersects the superior margin of the vertebral axis and the subject's hard palate. This plane is identified in the plotted data as pharyngeal position zero. Meanwhile, at the other end of the range, the image plane intersects the inferior margin of the sixth cervical vertebra and, depending on larynx height, either the glottis or the space immediately below it.

The imaging system used is an experimental whole body scanner developed at Yale University in collaboration with General Electric, employing a 0.15-T resis-

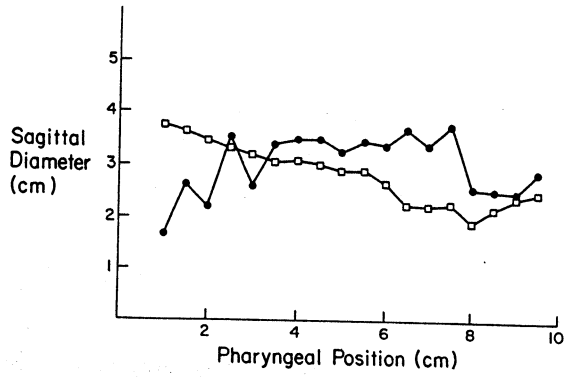


Fig. 3. Sagittal tract diameters plotted as a function of pharyngeal position for the vowel /i/ (see Method section for anatomical location of pharyngeal position zero). Rectangular points were measured from a xerographic x-ray and circular points were obtained from a series of MRIs.

spine is less curved and the mandible thrust further forward than is the case during the x-ray. Thus since the pharynx experiences less extension during an MRI, it is to be expected that larger cross-sectional areas might be observed as a result of the reduced elastic tension within the tissues of the pharyngeal wall.

Another possibility is that the discrepancy represents the normal variation in performance to be expected from two widely separate attempts to execute the same articulatory maneuver, notwithstanding the differences in posture. Thus both measures may be equally accurate representations of the articulatory configuration data to which they have been applied.

There is also a possibility that some error may result from our method of estimating the outline of the air–tissue boundary. This seems to be an unlikely explanation in general, since the boundaries are quite sharp over many regions of the vocal tract, and thus large changes in the threshold grey-level result in only small differences in estimated area. However, there may be larger errors in some regions where the image is noisy or has reduced dynamic range, especially near the extreme laryngeal end of the vocal tract where the rf receiver coil begins to rapidly lose sensitivity. Regions in which the area changes rapidly as a function of vertical position, or where the boundary may move during the 3.4 min required to obtain the image, are also likely to have less sharp boundaries and are more susceptible to measurement error based on the choice of threshold level.

An example of changes in pharyngeal cavity shape with vowel articulation is shown in Fig. 4. The upper image is for /i/ at about the level of the third cervical vertebra, as in Fig. 2. The lower image is for /a/ at the same level. A graph of the MRI-derived effective cross-sectional area as a function of pharyngeal

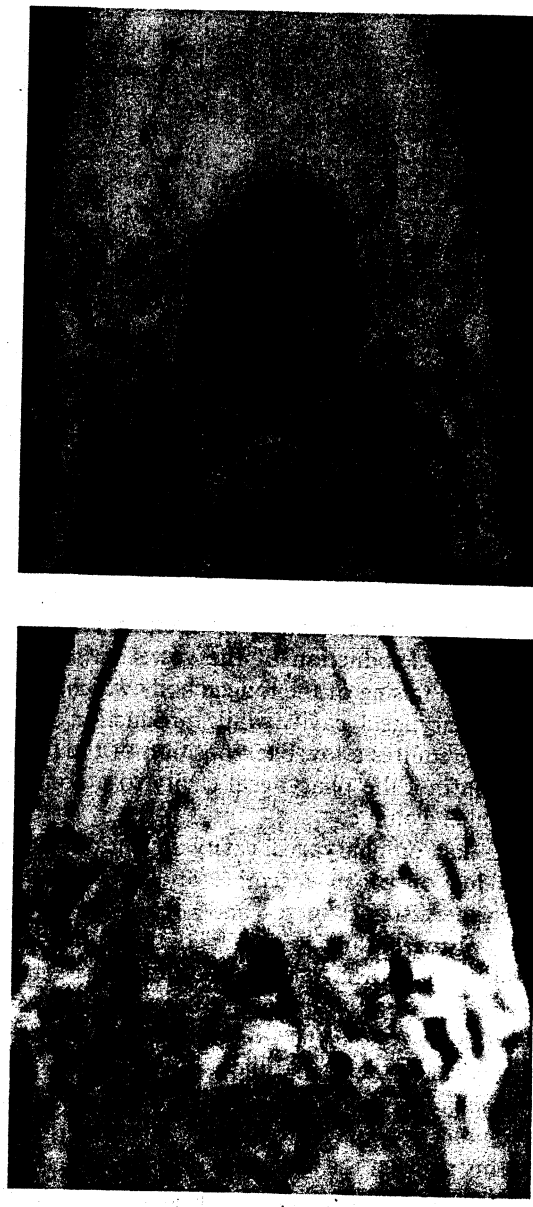

Fig. 4. MRI transaxial slices through the vocal tract during production of the vowels /i/ and /a/ by the same subject. The upper image is for the vowel /i/ and is an enlargement of the upper image in Fig. 2. The lower image is for the vowel /a/ and is taken at the same level as the upper image.

position (the pharyngeal area function) is shown plotted for the vowels /i/ and /a/ in Fig. 5. Linguists classify the former as a high front vowel—the term high front referring to the fact that, during articulation, the tongue body is raised and the highest point of the tongue surface occupies the anterior half of the buccal cavity. On the same basis, the latter vowel is classified as low back. Reflecting these extreme differences in production, it is perhaps not surprising that

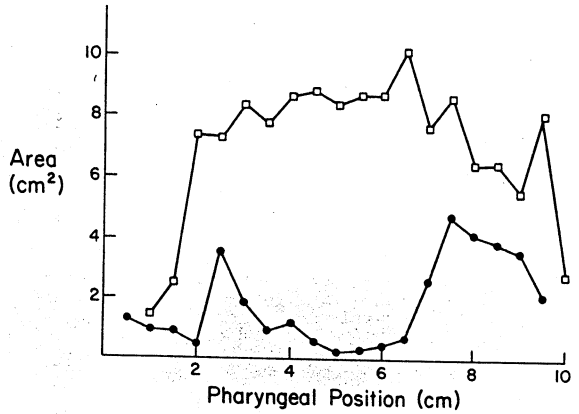


Fig. 5. Pharyngeal area functions plotted for the high front vowel /i/ (rectangular points) vs the low back vowel /a/ (circular points) that show the contrast in pharyngeal volume.

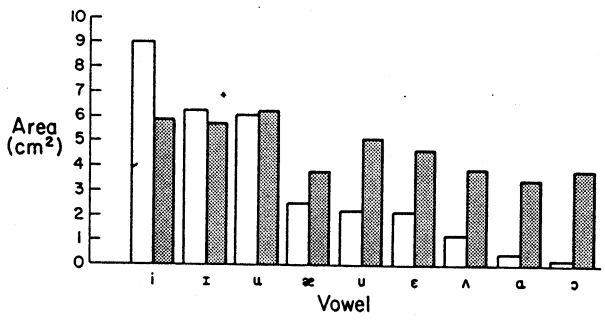


Fig. 6. Histogram of the cross-sectional areas at two positions in the pharynx (see Results section for anatomical location) during the production of an ensemble of vowels. Unfilled columns represent the cross-sectional areas at the uppermost position and strippled columns the areas at a point 3 cm below.

these vowels contrast strongly with respect to pharyngeal volume. Production of the vowel /i/ requires fronting and raising of the tongue body which creates a large pharyngeal cavity while production of the vowel /a/ requires a low back position of the tongue which confines the pharyngeal cavity to a considerably smaller volume.

Some data on pharyngeal cavity dimensions during the production of an ensemble of vowels that fall between the extremes of /a/ and /i/ are shown plotted as a histogram in Fig. 6. The data are the cross-sectional areas measured at two positions or planes of intersection with the pharynx: The uppermost plane intersected the tract at the level of the midpoint of the third cervical vertebra, and the second plane lay 3 cm below the first. The vowels in Fig. 6 have been ordered in descending cross-sectional area on the basis of measurements made at the upper level. At the lower level, the histogram reveals that the range of area change across the vowel ensemble is smaller than at the upper level. Moreover, the vowels /i/ and /I/ are characterized by larger areas in the upper pharynx than in the lower. With the exception of /u/, this feature distinguishes /i/ and /I/ from all the remaining vowels, for which the reverse is true. Finally, if each member of the ensemble is given its conventional linguistic classification where *A* and *P* represent the anterior and posterior tongue positions, respectively, and the digits 1–4 represent descending degrees of tongue height, a stepwise regression shows that tongue height is responsible for a majority 51% of the variance as compared to 14% for the anterior–posterior tongue motion. This result is consistent with observations from cine x-ray films which show that tongue raising results in an overall increase in the sagittal width of the pharyngeal cavity and a presumed increase in pharyngeal volume.

## CONCLUSIONS

MRI offers a promising noninvasive method of collecting data on the dimensions of the vocal tract during vowel production, particularly in the least-accessible pharyngeal regions. However, the work reported here has shown that further attention needs to be directed to the problem of increasing measurement reliability and accuracy. In one instance where independent verification has been sought, the tract dimensions obtained from MRIs have been significantly larger than those obtained from x-ray images. There are several possible sources for these discrepancies; differences in body posture, differences between successive productions of the same utterance and errors in image analysis. All of these problems promise to yield to further study. Therefore it cannot be long before the MRI technique will reach its full potential as the source of much needed data covering the full range of vocal tract configurations used by a single speaker.

## REFERENCES

1. Chiba, T.; Kajiyama, M. *The Vowel, Its Nature and Structure.* Tokyo: Tokyo-Kaiseikan; 1941.
2. Fant, G. *Acoustic Theory of Speech Production.* The Hague: Mouton; 1960.
3. Fant, G. Formants and cavities. E. Zwirner, W. Bethge eds. *Proceedings of the Fifth International Congress of Phonetic Sciences.* Basel; Karger; 1965: 120–140.
4. Gauffin, J.; Sundberg, J. Pharyngeal constrictions. *Phonetica* 35:157–168; 1978.
5. Heinz, J.M.; Stevens, K.N. On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *J. Acoust. Soc. Am.* 36(A):1037; 1964.
6. Johansson, C.; Sundberg, J.; Wilbrand, H.; Ytterbergh, C. From sagittal distance to area: A study of

transverse, cross-sectional area in the pharynx by means of computer tomography. *R. Inst. Technol. STL-QPSR* 4:39–49; 1983.

7. Kiritani, S.; Tateno, Y.; Iinuma, T.; Sawashima, M. Computed tomography of the vocal tract. M. Sawashima, F. Cooper eds. *Dynamic Aspects of Speech Production.* Tokyo: University of Tokyo Press; 1977: 203–206.
8. Ladefoged, P.; Anthony, J.F.K.; Riley, C. Direct measurement of the vocal tract. *UCLA Working Papers in Phonetics* 19:4–13; 1971. Abstract also in *J. Acoust. Soc. Am.* 49:104; 1971.
9. Mermelstein, P. Articulatory model for the study of speech production. *J. Acoust. Soc. Am.* 53:1070–1082; 1973.
10. Perkell, J.S. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* Cambridge: MIT Press; 1969.
11. Rubin, P.; Baer, T.; Mermelstein, P. An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.* 70:321–328; 1981.
12. Sundberg, J. On the problem of obtaining area functions from lateral x-ray pictures of the vocal tract. *R. Inst. Technol. STL-QPSR* 1:43–45; 1969.