AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION

1123

Predicting Midsagittal Pharynx Shape From Tongue Position During Vowel Production

D. H. Whalen
Haskins Laboratories
New Haven, CT

A. Min Kang
Haskins Laboratories
and
Yale University
New Haven, CT

Harriet S. Magen
Haskins Laboratories
New Haven, CT and
Rhode Island College
Providence

Robert K. Fulbright
John C. Gore
Yale University School of Medicine
New Haven, CT

The shape of the pharynx has a large effect on the acoustics of vowels, but direct measurement of this part of the vocal tract is difficult. The present study examines the efficacy of inferring midsagittal pharynx shape from the position of the tongue, which is much more amenable to measurement. Midsagittal magnetic resonance (MR) images were obtained for multiple repetitions of 11 static English vowels spoken by two subjects (one male and one female). From these, midsagittal widths were measured at approximately 3-mm intervals along the entire vocal tract. A regression analysis was then used to assess whether the pharyngeal widths could be predicted from the locations and width measurements for four positions on the tongue, namely, those likely to be the locations of a receiver coil for an electromagnetometer system. Predictability was quite high throughout the vocal tract (multiple $r > 0.9$), except for the extreme ends (i.e., larynx and lips) and small decreases for the male subject in the uvula region. The residuals from this analysis showed that the accuracy of predictions was generally quite high, with 89.2% of errors being less than 2 mm. The extremes of the vocal tract, where the resolution of the MRI was poorer, accounted for much of the error. For languages like English, which do not use advanced tongue root (ATR) distinctively, the midsagittal pharynx shape of static vowels can be predicted with high accuracy.

KEY WORDS: magnetic resonance imaging (MRI), vowel production, tongue, pharynx

It has been recognized since at least the 1920s that the shape of the pharynx greatly affects the acoustic realization of vowels (Huizinga, 1932; Russell, 1928). However, it is difficult to measure the shape in this portion of the vocal tract. Until quite recently, only conventional x-rays were capable of providing direct information about the size of the air space in the pharyngeal region. Unfortunately, this technique has limitations of interpretation and data collection. Specifically, three-dimensional structures are projected onto two-dimensional images. The consequence is that structures outside the plane of interest cast "shadows" in the image. This can make it difficult to focus on the midline of the tongue, where the width of the airspace is the most critical from an articulatory point of view. Furthermore, although the amount of radiation exposure is small, the length of time that any one subject can be exposed is limited; alternative methods that avoid such exposure are always preferable.

Although the early work showed the importance of the pharynx as well as the tongue, phonetic descriptions, for the most part, have continued to focus on the position of the tongue within the mouth (see

Ladefoged, 1967 for a review of the main trends). This stance is formalized in the description of vowels in the IPA as being based on "the three dimensions of height, frontness, and rounding" (International Phonetic Association, 1989). In addition to these dimensions, some languages use a feature called "advanced tongue root" (ATR) distinctively (Lindau, 1974; Pike, 1967; Stewart, 1967). Although this use of the pharyngeal space is recognized in the IPA (p. 74), the IPA makes no mention of the more common covariation of pharyngeal shape and tongue shape.

The development of magnetic resonance imaging (MRI) has made measurement of the pharynx in static shapes easier. MRI allows the safe study of vocal anatomy for static configurations, both in the midsagittal domain and in full volumetric measurements (Alwan, Narayanan, & Haker, 1997; Baer, Gore, Boyce, & Nye, 1987; Baer, Gore, Gracco, & Nye, 1991; Lakshminarayanan, Lee, & McCutcheon, 1991; Moore, 1992; Narayanan, Alwan, & Haker, 1995, 1997; Story, Titze, & Hoffman, 1996). MRI avoids the problem of shadows because of its ability to image a single plane of specified width and so can give more accurate results.

There are three main drawbacks to MRI. First, the subject must be in a supine position, and this can slightly alter the production of vowels (Tiede, Masaki, Wakumoto, & Vatikiotis-Bateson, 1997). Second, the imaging protocols used prevent the teeth from appearing in the images, necessitating their reconstruction (from casts or other techniques) for a full description of the vocal tract in this region. Third, acquisition times for high-quality MR images are slow, allowing only sustained phonations to be studied. (There has been some success recently in using rapid scans taken throughout different repetitions of a single utterance to make a composite movie [Masaki et al., 1997].)

Measurement of points on the tongue surface, on the other hand, has become much easier with the introduction of the x-ray microbeam system (Abbs & Nadler, 1987; Kiritani, 1986; Westbury, 1994) and electromagnetometer systems (Perkell et al., 1992; Schönle et al., 1987). These latter systems rely on the magnetic fields created by three magnets located around the speaker's head to allow the recovery of the positions of various receiver coils within that field. Because the magnetic field is unaffected by the composition of the vocal tract itself, these receivers can be located on the tongue, or even the soft palate, as well as the external articulators (lips and jaw). These systems collect data in real time and so can track speech as it is normally articulated.

The present work brings these two techniques together to make two determinations: first, whether descriptive linguistics' reliance on tongue position within the oral cavity as the description of vowel articulation is justifiable and, second, whether it will be possible to rely on magnetometer data for articulatory synthesis. On the first issue, we can be fairly certain that there will be some cross-linguistic limitation to the use of the body of the tongue alone for describing vowels because of those languages that use ATR distinctively (Lindau, 1974; Pike, 1967; Stewart, 1967). In these languages, differences in vowel category can be signaled by varying pharyngeal aperture while leaving the part of the tongue in the oral cavity in place (Ladefoged, 1964; Tiede, 1996). For the majority of languages, though, there does seem to be a high correlation between tongue position and pharyngeal shape, whether based on statistical modeling (Harshman, Ladefoged, & Goldstein, 1977; Nix, Papçun, Hogden, & Zlokarnik, 1996) or inferred from the success in predicting formant values from tongue shape alone (Fischer-Jørgensen, 1985). The present study will assess the consistency of this relationship and thus allow us to address the second issue: If the pharynx can be modeled based solely on measurements of the tongue surface, this would allow greater use of tongue body measurements to guide articulatory synthesis.

Method

Subjects

Two subjects participated. The male subject, M1, is a native of Maine and speaks a northeastern U.S. dialect that preserves the /ɑ/–/ɔ/ distinction. The female subject, W1, is a native of Illinois and speaks a midwestern U.S. dialect that also preserves the /ɑ/–/ɔ/ distinction. Both subjects are phonetically trained.

Stimuli

Eleven vowels of English were used: [i ɪ eʲ ɛ æ ɑ ʌ ɔ oʷ ʊ u]. Because vowel articulation is reasonably well represented by midsagittal shape alone (Mermelstein, 1973; Perrier, Boë, & Sock, 1992) and obtaining full volumetric data is much more time-consuming, only a small set of sagittal MR images (one midsagittal and one on either side) was obtained. Even so, acquisition of a single token required approximately 14 s. Therefore, only static vowels could be used. For the two vowels with large offglides, the subjects were instructed to hold the initial position throughout; compliance was monitored by listening over the intercom system during phonation.

Apparatus

MR images were obtained on a General Electric 1.5T Signa magnet. T1-weighted sagittal scans of the whole

head were obtained using a conventional spin-echo technique. Repetition time (TR) was 500 ms; echo time (TE) was 11 ms; imaging matrix was 256 × 192; the number of excitations (nex) was equal to 1. This prescan provided 5-mm contiguous slices of a 28-cm field of view (FOV) extending from the top of the head to about the level of the larynx. Given the drop-off in image quality at the periphery of the image, the larynx was not well represented.

From the prescan, three contiguous 5-mm slices straddling the midline were chosen for imaging. The simultaneous acquisition of three slices takes no longer than acquiring one using this protocol; the extra images were used to check for head movement during the experiment. This was done by verifying the presence of the nasal septum in the midsagittal slice only. (Any rotation or tilting of the head would have put the nasal septum out of the midsagittal plane.) T1-weighted images were obtained for each token using a fast spin-echo technique with the following parameters: TR = 400 ms, TE = 14 ms, echo train length (ETL) = 4, echo spacing (Esp) = 14 ms, 128 × 128, 1 nex, FOV = 28 cm.

Procedure

For each acquisition, subjects were prompted with the target vowel and its keyword (e.g. "The next vowel is /ɑ/ as in 'knock-kneed'") using the built-in intercom system. Two-syllable keywords were used for increased clarity because of the noisy conditions of the intercom system. The MRI scan was started as soon as the subject began phonation. Subjects sustained phonation until the end of the acquisition (approximately 15 to 16 s), which was signaled by the ceasing of noise generated by the MRI unit. Five tokens of each vowel were obtained for M1. Six tokens were obtained for W1 (except for the vowel /ɛ/, which had only four because of a processing error). Examples of midsagittal MR images for each of the 11 vowels for the male and female speakers are shown in Figures 1 and 2, respectively. None of the nonmidsagittal images indicated any noticeable movement of the head off of the midline.

Analysis

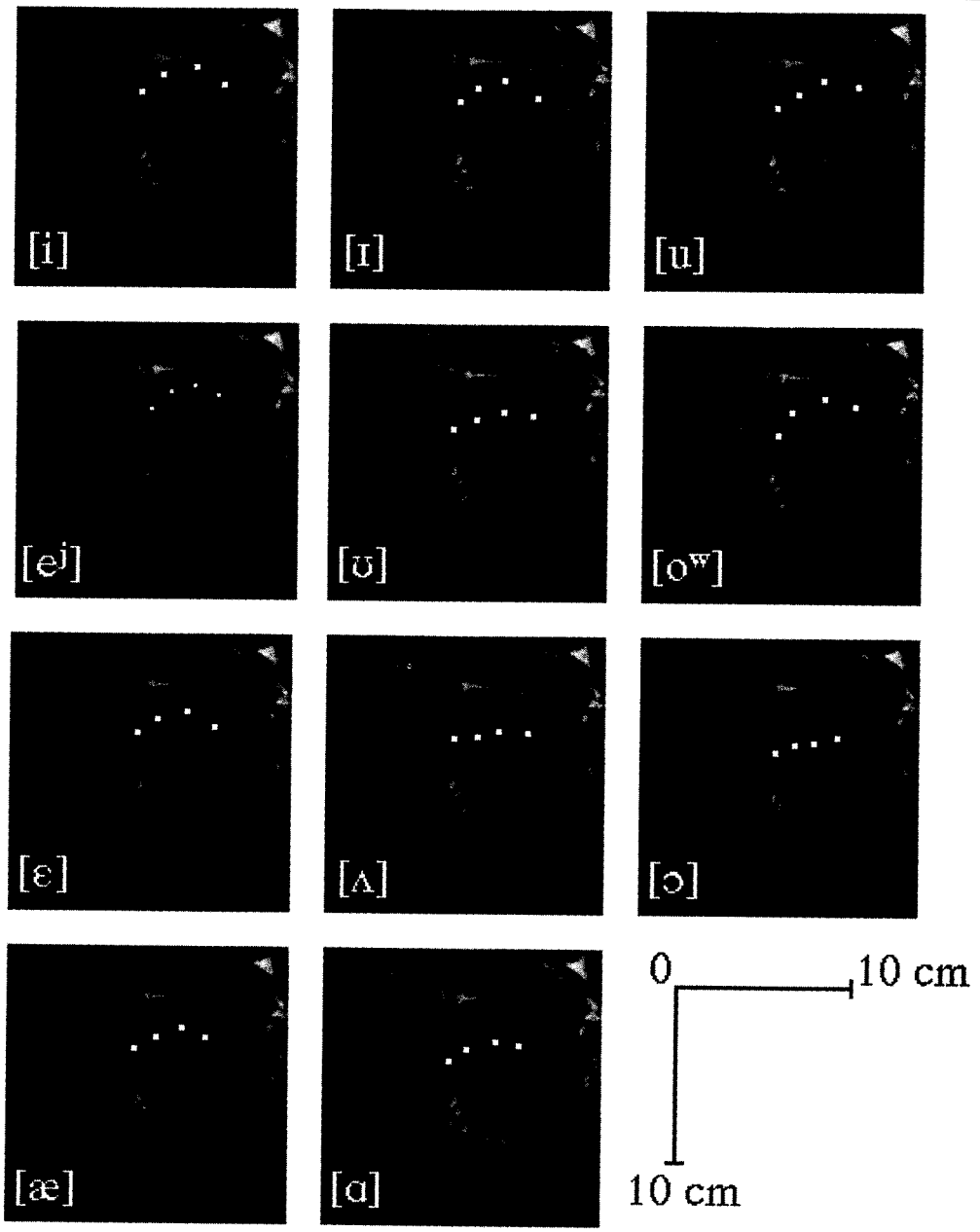
Vocal tract widths were measured from the MR images on a Power Macintosh 7300/180 computer using the public domain NIH Image v1.61 program (developed at the U.S. National Institutes of Health and available on the World Wide Web at http://rsb.info.nih.gov/nih-image/). Using macros originally developed by Mark K. Tiede of ATR Human Information Processing Research Labs that were modified to analyze sagittal images, four fixed reference points were selected on each image: alveolar ridge (P1), most superior point of the hard palate (P2), most superior and inferior points on the rear pharyngeal wall (P3 and P4, respectively). The placement of these points is illustrated in Figure 3(a).

These points were used to overlay a grid system; see Figure 3(b). In the pharyngeal region, parallel gridlines were placed 3 mm apart, perpendicular to an imaginary line joining P3 and P4. Gridlines in the curved uvular/velar region were plotted based on a circle fit to points P1, P2, and P3; gridlines radiated from the center of this circle every 5°. In the alveolar region, parallel gridlines were plotted perpendicular to an imaginary line joining P1 and P2—again every 3 mm. The center of curvature for the gridlines in the curved region around the teeth was the top of the gridline closest to P1. The parallel lines in the lip region were lined up with the edges of the image.

The width of the air space formed by the vocal tract at each gridline was measured automatically by searching for the air-tissue boundaries, indicated by abrupt changes in signal intensity (i.e., pixel values) along a gridline. Images were preprocessed to improve the detection. Because the uvula intruded into the oral cavity to different extents in different tokens, it was erased in each image by following the palatal contour. This eliminated the large increase in apparent depth between the gridline just before and the one just after the tip of the uvula. The inclusion or exclusion of the uvula does not seem to affect the acoustic output significantly in synthetic modeling (Magen & Kang, 1997). Lower in the pharynx, the epiglottis was defined as the edge of the airspace. The teeth are not present on the images, and thus the vocal tract width at these gridlines were gum-to-gum measurements. This area of the tract, however, was of only peripheral interest for the present experiment. All measurements were checked by the experimenters for any obviously large errors. For example, when running the algorithm on images of the first subject, M1, the nasal cavity (apparent above the soft palate in Figure 1) was often mistaken for the vocal tract, and the air-tissue boundary was frequently misidentified. To avoid these general errors when processing images from W1, the nasal cavity was filled in during the preprocessing and the air-tissue boundary refined within each image before application of the algorithm. Consequently, there were no measurements that needed correcting for W1.

On the images of each vowel, four positions on the tongue were selected as being the most likely locations for receiver coils to be placed during an electromagnetometer study. For W1, these estimations were made on the basis of our previous experience with this subject in such experiments and by referring to existing MR images in which pseudo-receivers (oil-filled capsules) were glued to the tongue (C. Browman & L. Goldstein,

Figure 1. Midsagittal MR images of the vowels examined for male speaker M1. One token (out of five collected) for each of the 11 vowels is shown. The four white squares on the tongue surface in each image represent the probable locations of magnetometer receivers. The receivers are designated R1 to R4, from anterior to posterior.

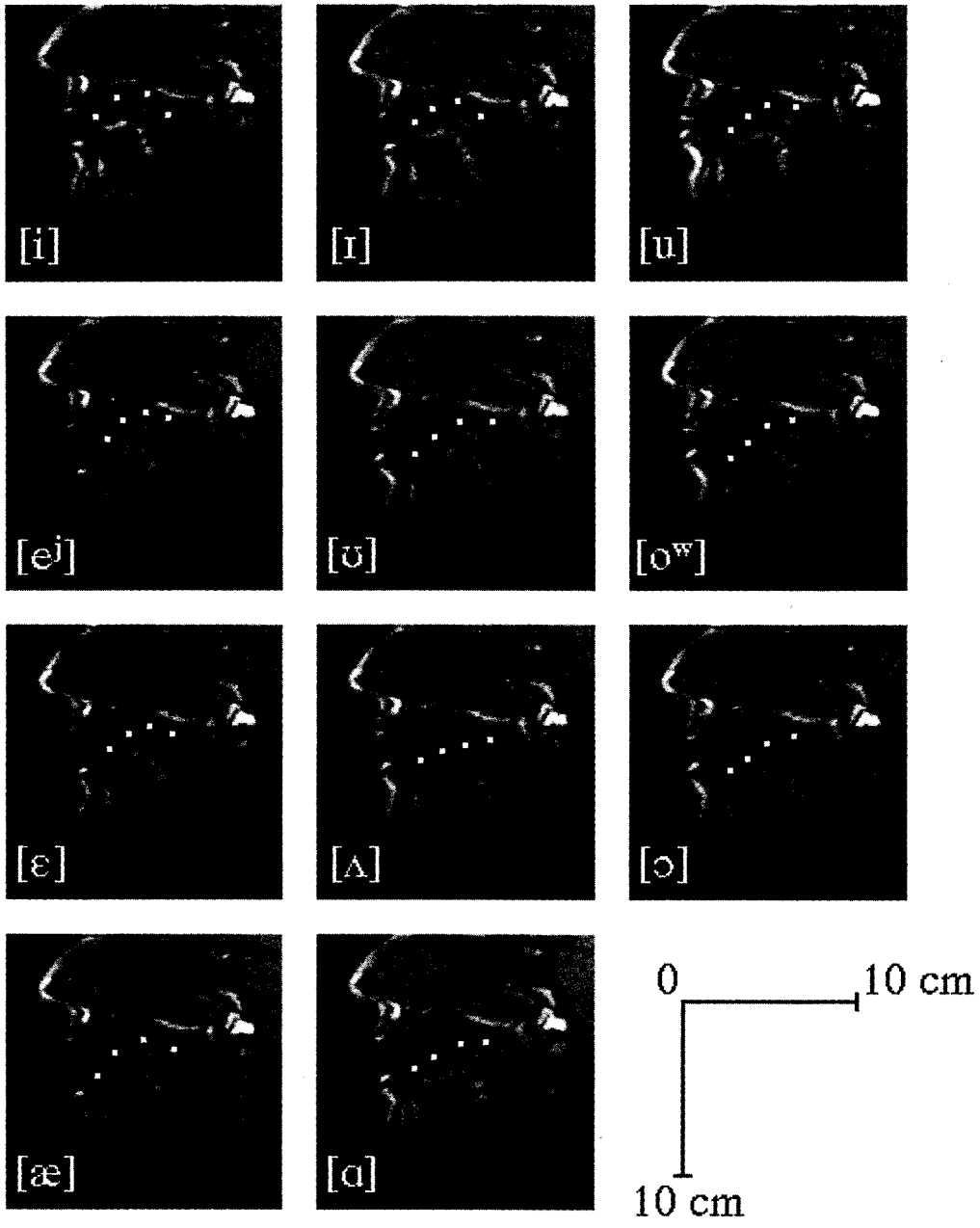


personal communication). (The pseudo-receivers were large enough to interfere with the size of the vowel cavity, especially for [i], and thus were not used in the present study.) For M1, the locations were approximated by comparison with the existing MR images of W1.

Our assessment of the predictability of the vocal tract widths was based on regression analyses. In these analyses, the independent variables consisted both of measurements of the MRI images and of linguistic characteristics of the vowels produced. The image-based variables consisted of two values associated with each of the four points reflecting the receiver positions: the location along the vocal tract (i.e., gridline) and the width at that location. The resulting eight values constituted the set of Measured variables. In addition, two Categorical variables (see Table 1) were defined on the basis of traditional values in descriptive linguistics: vowel height (five possible values: high, high-mid, mid, low-mid, and low) and vowel frontness (three values).

To predict the widths throughout the vocal tract,

Figure 2. Midsagittal MR images of the vowels examined for female speaker W1. Images are displayed as in Figure 1.

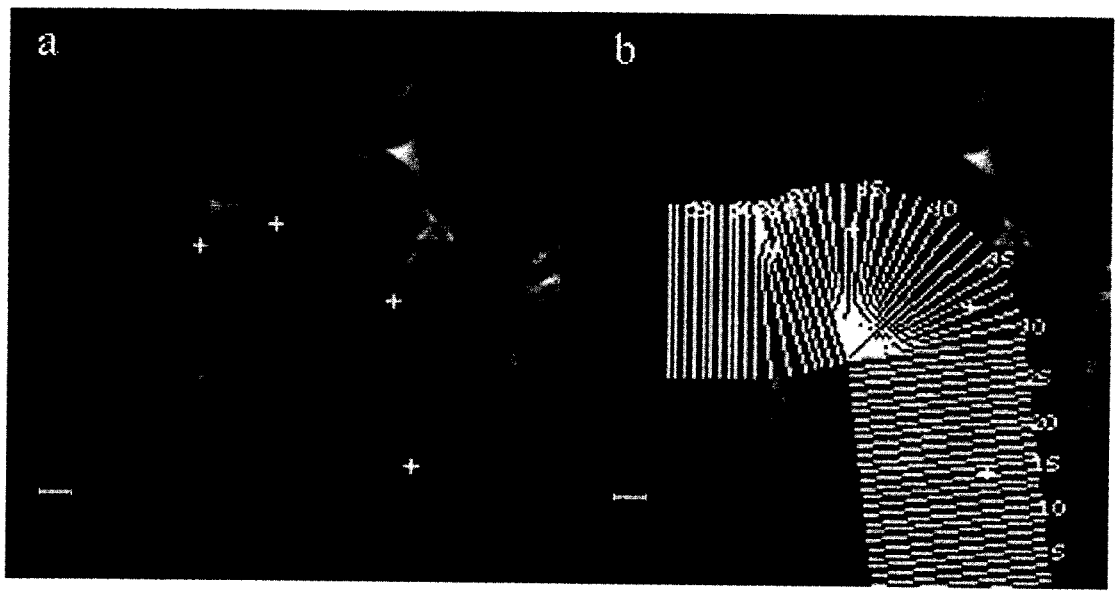


the above variables were used as independent variables in three different regression analyses: Measured, Categorical, and Combined. For all three analyses, the dependent measures were the widths of the airspace for each gridline, encompassing the individual tokens of all 11 vowels (55 for M1 and 66 for W1). An independent analysis was performed for each gridline (46 for M1 and 48 for W1). The first set of independent variables were the eight Measured variables based on the MR images. The Categorical analyses used only the two categorical vowel variables as predictors. The Combined analysis was also a multiple linear regression but used both subsets of variables (i.e., all 10 variables).

Results

Multiple r values for each gridline from the three regression analyses (Measured, Categorical, and Combined) for M1 and W1 are plotted in Figures 4 and 5, respectively. In general, the Measured and Combined analyses performed equally well ($r > 0.9$) over most of the vocal tract. All correlations were significant at, minimally, the

Figure 3. Midsagittal MR images of [a] produced by M1. (a) White crosses indicate the four fixed reference points used to overlay the grid system (P1-P4, labeled front-to-back). (b) Width measurements were taken using the superimposed gridlines. Parallel gridlines were placed 3 mm apart in the pharyngeal region, perpendicular to an imaginary line joining P3 and P4. Gridlines in the curved uvular/velar region were plotted based on a circle fit to points P1, P2, P3; gridlines radiated from the center of this circle every 5°. In the alveolar region, parallel gridlines were plotted perpendicular to an imaginary line joining P1 and P2, again every 3 mm. The center of curvature for the gridlines in the curved region around the teeth was the top of the gridline closest to P1. The parallel lines in the lip region were lined up with the edges of the image. Scale bar, 1 cm.



.05 level (most $ps < .00005$), with the exception of the most extreme gridline at the lips for M1. The analysis of variance that computes this significance corrects for any bias inherent in using a relatively small number of measurements (55 for M1 and 66 for W1) in the correlations. There was little gain for the Combined over the Measured. Only in the pharyngeal region for M1 was the Measured analysis not as effective as the Combined. The Categorical analysis was less effective throughout, as could be expected for an analysis using two independent variables as opposed to eight. It is striking, however, that predictability was as high as it was despite the use of only two nonmeasured variables.

Decreased performance for all analyses was found at the larynx and lips. Poor imaging in the laryngeal region was likely to be responsible. The lips also led to reduced image quality, most likely because of small movements during the production. Because an image is summed over time, these movements appear as blurring of the structures. Additionally, the definition of the relevant vocal tract width is problematic in this area. The small dips observed at gridline 25 for M1 and gridline 26 for W1 seem to be due to small changes in the location of the tip of the epiglottis. Changes in the functions specific to one talker were as follows: For M1, performance around the uvula was lower for all analyses and through much of the oral cavity for the Categorical analysis. For W1, there is a noticeable dip in performance in the rear oral region in the Categorical analysis alone.

Table 1. Values of the two categorical variables for the 11 English vowels. Higher vowels receive higher numbers on the Height dimension, whereas more front vowels receive lower numbers on the Frontness dimension.

Dimension	i	ɪ	eʲ	ɛ	æ	ɑ	ʌ	ɔ	oʷ	ʊ	u
Height	5	4	3	2	1	1	2	2	3	4	5
Frontness	1	1	1	1	1	2	2	3	3	3	3

To assess whether the effectiveness of the overall predictions would hold at the level of the individual vowel, the mean absolute residual from the Combined analysis was examined. The residuals consisted of the actual value for the depth of the air space minus the value predicted by the various regression coefficients for each token at each gridline. Combining all vowels and all gridlines, this was less than 1 mm for both subjects (.95 for M1 and .93 for W1). The percentage of individual tokens at which the mean absolute residual was less than two millimeters provides an overall measure of the effectiveness of the predictions: For M1 (the solid line in Figure 6), 88.8%; for W1 (the solid line in Figure 7), 89.5%. Errors for the pharyngeal region are higher

Figure 4. Results of the three correlation analyses for speaker M1. Multiple *r* (i.e., predictability) plotted for each gridline. The regions of the vocal tract corresponding to the gridlines are labeled along the abscissa. Open triangles show results from a multiple linear regression using eight Measured variables. Open circles show results from a multiple linear regression using only the two Categorical variables. Filled squares show a multiple linear regression combining both sets of variables (Measured + Categorical).

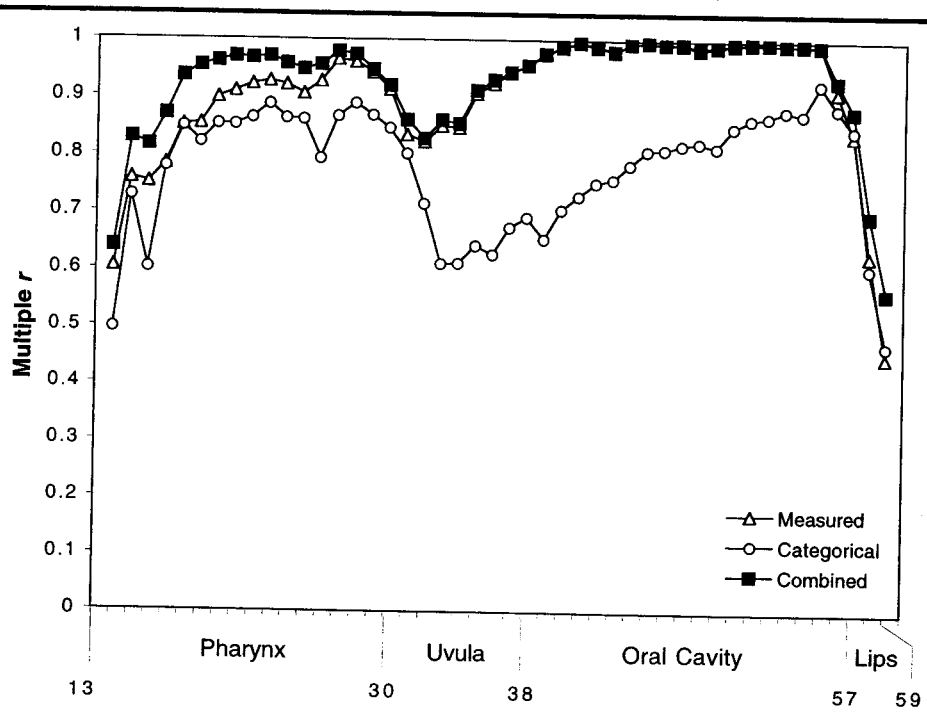


Figure 5. Results of the correlation analysis for speaker W1. The regression analyses are as in Figure 4.

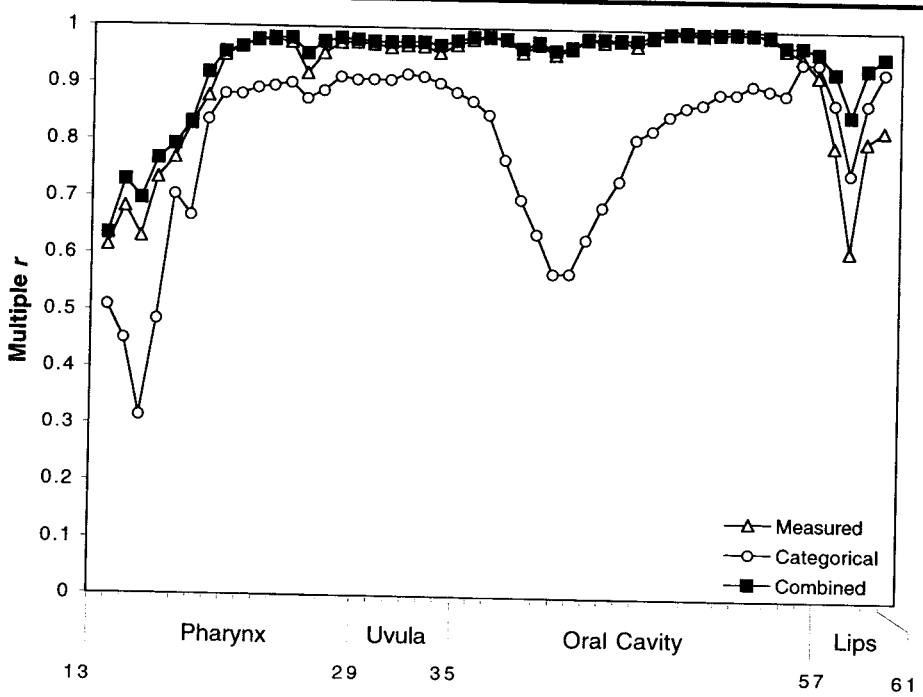


Figure 6. Absolute residuals from the Combined regression, speaker M1. The mean of all 11 vowels is shown in the solid line with filled symbols. Three selected vowels are shown with the dotted lines and hollow symbols. The value for the truncated /u/ function at gridline 57 is 5.66.

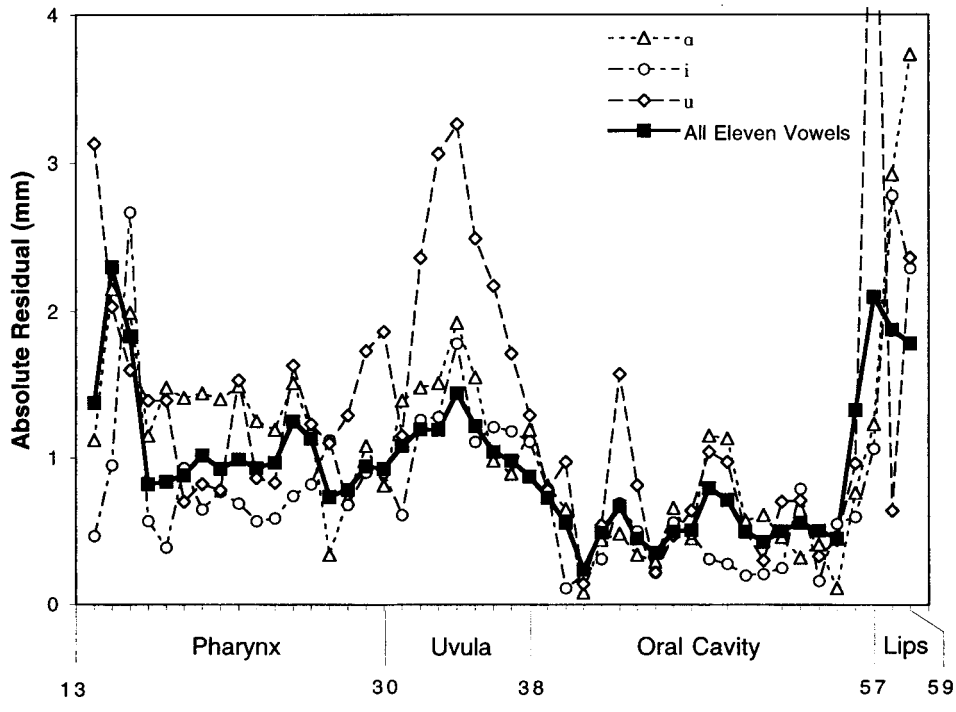
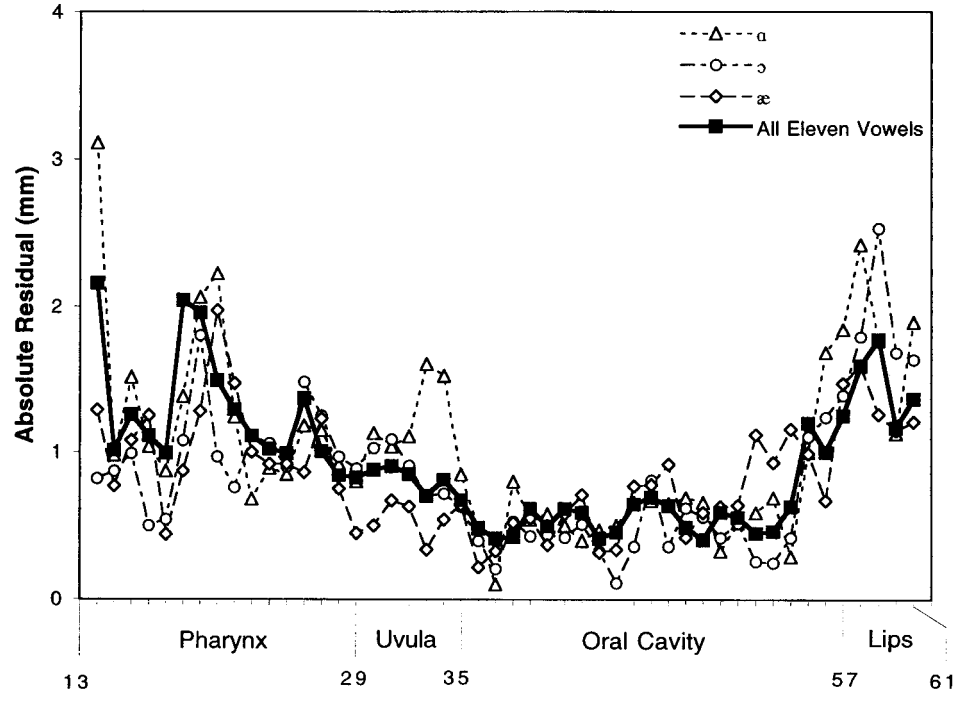


Figure 7. Absolute residuals from the Combined regression, speaker W1, shown as in Figure 6.



on average than those in the oral cavity (0.96 mm vs. 0.53 mm for M1, 1.18 mm vs. 0.62 mm for W1). Because direct measurements of the oral cavity were available to the correlation analysis, this is not surprising. The size of the error in both regions, though, is quite small.

Figures 6 and 7 also show means for three of the vowels separately. For M1, the point vowels /ɑ i u/ are shown. For W1, the low and back vowels /ɑ æ ɔ/ are shown. Although there is variability among the vowels in both sets, there is no obvious pattern to the differences. Examination of the remaining eight vowels also failed to show any one vowel that generated a unique pattern or any systematicity among vowels sharing phonetic features.

Inspection of the images for the two low vowels /ɑ/ and /ɔ/ revealed differences between the two subjects. Although the dialects of both speakers have a phonological difference between these two vowels, only the images for W1 showed any difference in the midsagittal MRI. It may be that M1 used only one of the vowels, despite prompting with keywords, or that he needed some context to maintain the distinction. We explored this by making acoustic recordings outside of the MRI unit. The conditions of our experiment were simulated as closely as possible, both by having the subjects supine and by having them sustain the vowels for about 14 s. Five repetitions of each of the 11 vowels were recorded. Formants were measured at 1-s intervals, giving about 70 measurements per vowel. The Appendix shows these values for both speakers, averaged across all the measurements. For both speakers there is a significant difference in acoustics between /ɑ/ and /ɔ/. A t test revealed a significant difference in both F1 and F2 for M1 [$t(145) = 18.56$ and 19.19 respectively, $p < .0001$] and W1 [$t(129) = 7.34$ and 5.03, $p < .0001$]. Therefore, it is likely that the two speakers implement the /ɑ/–/ɔ/ distinction differently and that only the strategy used by W1 is apparent in the midsagittal plane. Speaker M1 may also rely on lip posture, which is not as reliably seen in the MR images. Alternatively, there may be changes in articulation off the midline that are important for this distinction. A full, volumetric study of M1's vowels would help in resolving this issue.

Discussion

Measurements made on magnetic resonance images of the vocal tract during sustained vowels were used to assess two points: (1) the adequacy of linguistic descriptions of vowels that make use only of tongue features and ignore the pharynx and (2) the likelihood that measurements of the tongue alone would be sufficient to guide the modeling of the pharynx for articulatory synthesis of speech. Images were collected for two speakers, one male and one female. High correlations for both speakers were obtained for analyses using eight Measured variables (derived from four locations on the tongue), two Categorical variables (the linguistic categories of vowel height and frontness), and the Combined analysis using all 10 variables. Although residual errors were larger in the pharyngeal region, the absolute magnitude of the errors was small. In the Combined analysis, errors of 2 mm or less were found in 88.8% of the cases for the male speaker and 89.5% for the female speaker. The Measured analysis was slightly less accurate, with 84.4% of the cases being 2 mm or less for both subjects. The Categorical analysis was noticeably less successful, with 55.3% of these small errors for M1, 62.1% for W1. This is still a substantial amount of the data, however. For the vowels studied, then, both the linguistic description and the application to articulatory synthesis are supported.

The adequacy of the Categorical model is impressive considering that it contains only two independent variables and takes no account of token-to-token variation. In the Combined analysis, the errors obtained are of approximately the same magnitude as our measurement resolution in the MR images (about 1 mm). This result indicates that there is substantial variability in tongue position for vowels that is nonetheless correlated with variability in the shape of the pharynx. We might expect that variability in pharynx shape would not correlate well with variability in tongue shape, because speech articulators typically are more tightly controlled for the important constriction and more variable in the remainder of the tract. Such increases in variability at nonessential locations have been found in the production of vowels (Beckman et al., 1995) and stop consonants (Papçun et al., 1992). The improvement of the Combined analysis over the Categorical indicates that variability in the oral cavity was correlated with that in the pharynx. Whether or not this indicates speaker-dependent amounts and regions of variation (Johnson, Ladefoged, & Lindau, 1993), it does reinforce the notion that the acoustically important pharynx is redundantly specified in the tongue position, even on a token-to-token basis.

For a language like English, then, the predictability of the pharyngeal widths means that using only descriptions of the front part of the tongue will be sufficient to differentiate the vowels. The IPA reliance on tongue position in the oral cavity for defining the vowel symbols, then, can be considered roughly accurate for such languages. From a phonetic point of view, however, it may be that such descriptors can make certain processes difficult to explain if the area of greatest constriction is indeed what is most important. For example, the tendency of pharyngeal consonants to induce "low" articulation in vowels might seem less mysterious if the vowels were simply described according to their narrowest, or

pharyngeal, constriction (McCarthy, 1994).

Although the description of vowels by tongue position is thus sufficient for predicting the pharynx shape for most languages, it should fail for languages using advanced-tongue-root (ATR) distinctively. Such languages contrast two vowels with the same tongue position but with different pharyngeal settings. For example, MRI data for Akan show virtually no difference in the tongue shape for plus and minus ATR vowels, yet the pharyngeal shape is vastly different (Tiede, 1996). Although this justifies the use of the feature ATR as the relevant one, it also indicates that languages such as Akan will not be well modeled by magnetometer measurements that do not include the pharynx.

For all analyses, two regions with decreased, though still significant, predictability were the larynx and the lips, but both of these declines are more likely due either to imaging errors or definitional problems. The low values for the lower pharynx, near the larynx, are likely to result from the measurements themselves, because of a drop-off in image resolution. This could be resolved by using an imaging system that could keep this portion of the vocal tract in the better-defined area. Such a set-up was not available for the present experiment. With the lips, there is a problem in defining the vocal tract itself. The flaring of the tube results in some uncertainty about what constitutes the end of the vocal tract (Story et al., 1996). Thus it is to be expected that these measurements will be less accurately modeled, because even examination of individual tokens does not always make it apparent which is the exact measurement needed. Better imaging resolution should help with the predictability of the lower pharynx, but the lip values may continue to be problematic.

The one other area of the vocal tract that resulted in lower correlations was the uvula of the male speaker. This seems to represent a region with more genuine variability that is not well predicted by the factors in the analysis. As can be seen in Figures 1 and 2, and especially Figure 1, there is a good deal of variability in the degree to which the uvula extends into the vocal tract's air space for each vowel. This variation was present across tokens of a single vowel for speaker M1. To simplify the analysis, the intrusion of the uvula into the oral cavity was erased from the images for both speakers by following the contour of the palate. However, all analyses still had some difficulty in this region for M1, suggesting a change in the contour in this region correlated with the uvular differences.

The present results allow us to consider using easily obtained measurements of the tongue to guide articulatory synthesis. Such systems typically use a two-dimensional representation of the sagittal vocal tract as input to a series of formulas for translating those dimensions into three-dimensional tubes whose acoustic output can be generated. This is specifically the case for the Haskins Articulatory Synthesizer (ASY; Mermelstein, 1973; Rubin, Baer, & Mermelstein, 1981). With the equations obtained from the present correlations, tongue locations obtained by the magnetometer (Perkell et al., 1992) can be used to generate estimates of the entire midsagittal vocal tract. It might, at some point, also be possible to do the same with the x-ray microbeam data archive (Westbury, 1994), depending on future developments in making the predictions speaker-independent.

One remaining concern about such synthesis, however, is that small changes in articulator position can result in large changes in acoustic output. This is most clearly the case when the tongue is near enough to the upper vocal tract to generate noise. If that distance is changed by as little as 2 mm, the result could either be a stop or a vowel, depending on whether the constriction is further narrowed or, instead, widened. However, the present results are aimed primarily at vowel articulations, and these are less sensitive to changes of such magnitude. Changes of 2 mm in pharynx width were modeled with ASY for shapes typical of the point vowels /ɑ i u/. The change in F1 that resulted was 24, 24, and 9 Hz, respectively. These represent a 3–7% shift in F1, a change too small to be linguistically significant (or perhaps even detectable; Mermelstein, 1978). Because ASY does not currently contain an aerodynamic component, it appears that the magnitude of errors that we obtain with the present correlations will be relatively unimportant in our synthesis projects.

Another potential limitation on the prediction of the pharynx shape is that coarticulation during running speech may further change the relationship between tongue shape and pharynx shape. The present study used vowels sustained for approximately 50 times as long as a vowel in running speech; this presumably alters the articulation. However, the more important issue is whether the co-occurrence of consonants and other syllables affects the relationship between the tongue shape and the pharynx shape. It may be that the relationship is unchanged, so that magnetometer measurements, for example, will be accurate predictors of pharynx shape even during running speech. But even if this relationship is unchanged, coarticulation can affect articulation of the intended vowel enough to make it substantially different from static versions. Other techniques, such as MRI movies (Masaki et al., 1997), may eventually allow us to address the coarticulation question.

The formulas that translate midsagittal distance into cross-sectional area may need to be improved for articulatory synthesis. The lack of a noticeable difference in the midsagittal shape for /ɑ/ and /ɔ/ for M1 may

indicate that this vowel uses modifications off the midline. Tiede's results for Akan suggest that the midsagittal distance to area cross-section function must be language-sensitive as well, because there were differences between English and Akan not only in the width but in the shape of the pharyngeal cavity (Tiede, 1996). These possibilities are best explored with full volumetric measurements of the vocal tract.

These present results should prove useful for the synthesis of speech from articulatory data, a research goal of this laboratory. Although the algorithms for such synthesis have existed for some time (Coker, 1967; Mermelstein, 1973; Rubin et al., 1981), the extraction of real articulatory data on which to base the models for the synthesis has lagged. The present study shows that a set of predictors from measurements of the tongue (e.g., with a magnetometer), combined with MRI data, can be used to approximate widths along the entire vocal tract, even the pharyngeal region. Related work has shown that matching articulatory synthesizer parameters to these MR images often results in accurate acoustic matches, although not for all of the vowels (Magen & Kang, 1997). Future studies will be needed to assess whether a speaker-independent set of predictors can be generated, but the ability to use a small set of MRI measurements in conjunction with easily obtained, dynamic magnetometer data is of great assistance to the generation of articulatory synthesis.

Acknowledgments

This research was supported by NIH Grant DC-02717 to Haskins Laboratories. We are grateful to Hedy Sarofin and Terry Hickey for technical assistance. Portions of this research were presented at the 134th Meeting of the Acoustical Society of America, December 1997, in San Diego, CA. We thank Alice Faber, Bryan Gick, Maureen Stone, John Hogden, and two anonymous reviewers for helpful comments on previous versions of this paper.

References

Abbs, J. H., & Nadler, R. D. (1987). *User's manual for the University of Wisconsin x-ray microbeam*. Madison, WI: Waisman Center.

Alwan, A. A., Narayanan, S. S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data. Part II: The rhotics. *Journal of the Acoustical Society of America, 101,* 1078–1089.

Baer, T., Gore, J. C., Boyce, S. E., & Nye, P. W. (1987). Application of MRI to the analysis of speech production. *Magnetic Resonance Imaging, 5,* 1–7.

Baer, T., Gore, J. C., Gracco, L. C., & Nye, P. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America, 90,* 799–828.

Beckman, M. E., Jung, T.-P., Lee, S., de Jong, K., Krishnamurthy, A. K., Ahalt, S. C., Cohen, K. B., & Collins, M. J. (1995). Variability in the production of quantal vowels revisited. *Journal of the Acoustical Society of America, 97,* 471–490.

Coker, C. H. (1967). Synthesis by rule from articulatory parameters. In *Proceedings of the 1967 Conference on Speech Communication and Processing* (Paper A9, pp. 52-53). New York: IEEE.

Fischer-Jørgensen, E. (1985). Some basic vowel features, their articulatory correlates, and their explanatory power in phonology. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 79–99). New York: Academic Press.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *Journal of the Acoustical Society of America, 62,* 693–707.

Huizinga, E. (1932). Über die Stelle, wo der Charakter des Selbstlautes gebildet wird. *Archives Néerlandises de Phonétique Experimentale, 7,* 104–117.

International Phonetic Association. (1989). Report on the 1989 Kiel Convention. *Journal of the International Phonetic Association, 19,* 67–80.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America, 94,* 701–714.

Kiritani, S. (1986). X-ray microbeam method for the measurement of articulatory dynamics: Techniques and results. *Speech Communication, 45,* 119–140.

Ladefoged, P. (1964). *A phonetic study of West African languages: An auditory-instrumental survey*. Cambridge, U.K.: Cambridge University Press.

Ladefoged, P. (1967). The nature of vowel quality. In *Three areas of experimental phonetics* (pp. 50–142). London: Oxford University Press.

Lakshminarayanan, A. V., Lee, S., & McCutcheon, M. J. (1991). MR imaging of the vocal tract during vowel production. *Journal of Magnetic Resonance Imaging, 1,* 71–76.

Lindau, M. (1974). The feature Advanced Tongue Root. In E. Voeltz (Ed.), *Third Annual Conference on African Linguistics* (pp. 27–34). Bloomington: Indiana University Publications.

Magen, H. S., & Kang, A. M. (1997). Effects of the uvula and the epiglottis on measurements of vowel production: Assessment by synthesis. *Journal of the Acoustical Society of America, 102,* 3093(A).

Masaki, S., Tiede, M. K., Honda, K., Shimada, Y., Fujimoto, I., Nakamura, Y., & Ninomiya, N. (1997). MRI observation of dynamic articulatory movements using a synchronized sampling method. *Journal of the Acoustical Society of America, 102,* 3166(A).

McCarthy, J. J. (1994). The phonetics and phonology of Semitic pharyngeals. In P. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 191–233). Cambridge, U.K.: Cambridge University Press.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53,* 1070-1082.

Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *Journal of the Acoustical Society of America, 63,* 572–580.

Moore, C. A. (1992). The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images. *Journal of Speech and Hearing Research, 35,* 1009–1023.

Narayanan, S. S., Alwan, A. A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *Journal of the Acoustical Society of America, 98,* 1325–1347.

Narayanan, S. S., Alwan, A. A., & Haker, K. (1997). Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data. Part I: The laterals. *Journal of the Acoustical Society of America, 101,* 1064–1077.

Nix, D. A., Papçun, G., Hogden, J., & Zlokarnik, I. (1996). Two cross-linguistic factors underlying tongue shapes for vowels. *Journal of the Acoustical Society of America, 99,* 3707–3717.

Papçun, G. J., Hochburg, J., Thomas, T. R., Laroche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the Acoustical Society of America, 92,* 688–700.

Perkell, J. S., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America, 92,* 3078–3096.

Perrier, P., Boë, L.-J., & Sock, R. (1992). Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients. *Journal of Speech and Hearing Research, 35,* 53–67.

Pike, K. L. (1967). Tongue-root position in practical phonetics. *Phonetica, 17,* 129–140.

Rubin, P. E., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70,* 321–328.

Russell, G. O. (1928). *The vowel: Its physiological mechanism as shown by x-ray.* Columbus: Ohio State University Press.

Schönle, P., Grabe, K., Wenig, P., Hohne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language, 31,* 26–35.

Stewart, J. M. (1967). Tongue root position in Akan vowel harmony. *Phonetica, 16,* 185–204.

Story, B. H., Titze, I. R., & Hoffman, E. A. (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America, 100,* 537–554.

Tiede, M. K. (1996). An MRI-based study of pharyngeal volume contrasts in Akan and English. *Journal of Phonetics, 24,* 399–421.

Tiede, M. K., Masaki, S., Wakumoto, M., & Vatikiotis-Bateson, E. (1997). Magnetometer observation of articulation in sitting and supine conditions. *Journal of the Acoustical Society of America, 102,* 3166(A).

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook.* Madison, WI: Waisman Center, University of Wisconsin.

Received April 8, 1998

Accepted December 14, 1998

Contact author: Doug Whalen, PhD, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511. Email: whalen@haskins.yale.edu

Appendix. Formant values (in Hz) obtained outside the MRI unit for speaker M1 (Part A) and W1 (Part B)

A.

Vowel	F1	F2	F3
ɑ	688	1138	2800
æ	616	1703	2673
ɔ	615	1012	2826
ʌ	596	1221	2885
ɛ	532	1739	2771
ʊ	461	1114	2833
e^j	411	2093	2700
ɪ	409	1968	2765
o^w	378	830	2780
u	255	845	2480
i	242	2510	2950

B.

Vowel	F1	F2	F3
ɑ	873	1378	3026
æ	820	1790	2552
ɔ	763	1254	2506
ʌ	785	1441	2807
ɛ	707	1881	2670
ʊ	439	2211	3292
e^j	425	2059	2531
ɪ	389	2447	3104
o^w	461	975	2334
u	483	1285	2499
i	368	947	2532