Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data

Andrew J. Lundberg[a)]
Department of Computer Science, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218

Maureen Stone
Division of Otolaryngology, University of Maryland Medical School, 16 South Eutaw Street, Suite 500, Baltimore, Maryland 21201

(Received 8 September 1998; revised 10 June 1999; accepted 29 June 1999)

This paper discusses methods for reconstructing the tongue from sparse data sets. Sixty ultrasound slices already have been used to reconstruct three-dimensional (3D) tongue surface shapes [Stone and Lundberg, J. Acoust. Soc. Am. **99**, 3728–3737 (1996)]. To reconstruct 3D surfaces, particularly in motion, collecting 60 slices would be impractical, and possibly unnecessary. The goal of this study was to select a sparse set of slices that would best reconstruct the 18 measured speech sounds. First a coronal sparse set was calculated from 3D surface reconstructions. Selection of contours was globally optimized using coarse to fine search. Sparse and dense reconstructions were compared using maximum error, standard deviation error, and surface coverage. For all speech sounds, maximum error was less than 1.5 mm, standard deviation error was less than 0.32 mm, and average reconstruction coverage was 80%. To generalize the method across subjects, optimal slice locations were calculated from only the midsagittal contour. Six midsagittal points were optimized to reconstruct the midsagittal contour. Corresponding coronal slices were then used to reconstruct 3D surfaces. For data collection planning, a midsagittal sample can be collected first and optimal coronal slices can be determined from it. Errors and reconstruction coverage from the midsagittal source set were comparable to the optimized coronal sparse set. These sparse surfaces reconstructed static 3D surfaces, and should be usable for motion sequences as well. [S0001-4966(99)03610-3]

PACS numbers: 43.70.Jt [AL]

INTRODUCTION

Ultrasound imaging has been used to represent tongue positions for over 15 years (Sonies *et al.*, 1981; Keller and Ostry, 1983; Stone *et al.*, 1983). Like other imaging systems, it provides a 2D measurement of the tongue surface contour in a single plane (such as midsagittal, coronal, or oblique). One strength of ultrasound is that it images tongue contour movement using a fairly rapid frame rate (30 Hz). Another is that contours from several spatial planes can be reconstructed into 3D surfaces (Stone and Lundberg, 1996). However, as ultrasound collects 2D slices, the subject must repeat the speech corpus once for each desired slice. Also, both contours and surfaces are represented by many points. Thus, compact quantification of movement is difficult in the case of contours and even more difficult in the case of surfaces. Two improvements of the ultrasound technique would radically increase its usefulness: reduced dimensionality of a tongue contour or surface, and accurate representation of surface motion. Prior research has accurately represented static 3D tongue surface shapes from dense data sets of ultrasound images (Stone and Lundberg, 1996). The present paper presents methodology to accurately represent static 3D tongue surface shapes and motion from sparse data sets of ultrasound images. The method also reduces the dimensionality of tongue surface representation and maintains highly accurate reproduction of local deformation features. This modification is an essential step if multi-plane tongue movements are to be reconstructed practically into tongue surface movements. Ultrasound has been shown to be a useful tool for collecting 3D tongue surface data. It is noninvasive, has no exposure limits, and is relatively inexpensive ($20 000 USD).

In previous research, a series of 2D images was used for reconstructing a detailed 3D view of the tongue surface (Stone and Lundberg, 1996). This required a special transducer, however, which collected 60 slices in a polar sweep of 60 degrees in 10 s. While this was feasible for a 3D static speech sample, this method is too slow for 4D data collection (3D surfaces moving in time). As there are not yet any 3D ultrasound devices that simultaneously collect multiple 2D slices in motion, any 4D sample would need to be repeated N times (where N is the number of slices to be used in the reconstruction).

The motion of three-dimensional tongue surfaces is of interest because the tongue is a complex system that is critical in speaking, swallowing, and breathing. The tongue is a volume preserving, deformable object. That is, tongue shape is systematically related to tongue position, because tongue volume can be redistributed, but not increased or decreased. Complicated tongue surface shapes can be produced in 2D and 3D due to the complex distribution of the tongue's

a)Electronic mail: lundberg@cs.jhu.edu

muscles and its lack of bony tissue (Kier and Smith, 1985). The intrinsic muscles originate and insert on soft tissue; the extrinsic muscles also insert on soft tissue, thus insuring deformation with every movement. Contracting various muscles and contacting the palate allows complicated shapes to be made. Subtle changes in these shapes reflect difference due to coarticulation (Stone and Lele, 1992), dialect and language (Stone and Yeni-Komshian, 1991), and speech disorders (Stone, 1995). In order to capture subtle shape changes due to these factors, especially over time, accurate tongue surface representation is essential.

The present study used the database acquired in Stone and Lundberg (1996) to determine a minimal number, or sparse set, of coronal slices needed for reasonable reconstructions. The specific coronal slices to collect must be specified, as well as what error tolerances define a reasonable reconstruction. Earlier work (Miyawaki *et al.*, 1975; Stone, 1990) suggested that 3D tongue surface shape could be adequately specified using five lengthwise segments. Therefore, although many sparse sets of coronal slices were tested, five slices were hypothesized to be optimal. In fact, six slices were determined to give the most accurate compact representation, as discussed below.

While real-time 3D ultrasound devices do not yet exist, there are experimental ultrasound systems that simplify the data collection of multiple static 2D slices. These approaches use a 2D ultrasound transducer with an automated 3D spatial positioning system. One system is the scanning 3D transducer used in Stone and Lundberg (1996), which internally moves a 2D transducer through a radial space (Acoustic Imaging Inc., Phoenix, AZ). The second is a holder that externally rotates a traditional 2D transducer (Tomtec Inc., Denver, CO). The first is static; the second allows time-varying data to be collected independently at several slices and then reconstructed. Both systems collect up to 60 planes of data and use computer control to position the transducer, but the commercial reconstruction algorithms are quite poor, and slices in planes, other than the original, are very unrealistic. Moreover, measurement of 60 tongue planes at 30 frames per second is unrealistically time consuming and unnecessarily dense spatially.

An alternative method, data-driven slice selection, calculates from subject data an optimal sparse source set of coronal slices from which reasonable 3D surfaces can be constructed. For this method, an externally rotated transducer, or even a manually positioned system, would be sufficient for 4D data collection. Thus, to collect data for 4D reconstructions, one would do multiple data collections of coronal tongue image sequences at a few specific orientations. The separate image sequences would then be aligned in 3D space from their respective collection orientations. In this paper, data-driven slice selection is simulated by selecting a sparse set of slices from the already existing dense set of coronal slices described in Stone and Lundberg (1996).

I. METHODS

A. Subject and speech materials

The subject was a 26-year-old white female with a Baltimore Maryland accent. Nineteen English speech sounds were studied: /i/, /ɪ/, /e/, /ɛ/ /æ/, /ɑ/, /ʌ/, /ɔ/, /o/, /ʊ/, /u/, /ɝ/, /ɜ/, /θ/, /ʃ/, /s/, /l/, /n/, /ŋ/. The subject sustained each phoneme for 10 s while 60 slices were collected. The data was collected with the subject lying on her back because the 3D transducer required a non-upright position for clear images (due to a fluid bubble within the transducer head that would obscure ultrasound when the transducer was in an upright position). Complete recording procedures and subject information can be found in Stone and Lundberg (1996). In addition, for this experiment, several additional tongue surfaces were collected. For validation of the reconstruction procedures the vowel /æ/ was collected twice using sagittal and coronal slices. To study variability the sound /l/ was collected twice, once with a normal articulation, and once pushing the tongue tip into the palate to induce variability.

B. Validation of dense data set

Prior to determining an optimal sparse source set for the 3D reconstructions, validation of the dense data sets was performed to guarantee the accuracy of the original (dense) data reconstructions. Further validation was performed on the reconstruction method beyond the original phantom reconstruction of a known surface done in Stone and Lundberg (1996). Data collection and reconstructions of the tongue surface were done from the coronal and sagittal dense data sets for /æ/. The first collection was done with the 60 slices oriented in the coronal direction. For the second data set, the transducer was rotated 90 degrees so that the 60 slices were oriented in the sagittal direction. Surface reconstruction was performed on each data set, and they were overlaid in 3D space to find the best fit in terms of overlap and minimal error (measured as the 3D distance between surface points on the data sets). It should be noted that this error would include the normal variability between repeated tokens of the same phoneme. For the coronal and sagittal /æ/ data, the maximum error achieved was 2.6 mm, with a standard deviation error of 1.16 mm, and 86% overlap of the two data sets (see Fig. 1). Figure 1 shows the reconstructed /æ/ surfaces from coronal (left) and sagittal (right) slice sets and a set of distance vectors (bottom) comparing the two surfaces. The distance vector image is a set of vectors from the first (left) reconstructed surface to the second (right) surface. The length of any vector corresponds to the 3D distance between the surfaces at that point (the orientation of the vectors is not necessarily in the direction of the shortest 3D distance between the surfaces).

These intersurface differences were largely due to differences in measurement errors that occur in coronal and sagittal tongue contours when using ultrasound. Tissue edges become difficult to measure whenever the surface is oblique to the ultrasound beam. This is most problematic for sagittal images when the tongue surface is grooved, and a sagittal contour may lie entirely along the descending slope of the groove. In the coronal plane, this is most problematic in the tongue root, where the entire contour may be oblique to the ultrasound beam. In general, one can recover a groove more easily from a coronal scan, and one can measure more of the tongue root on sagittal scans.

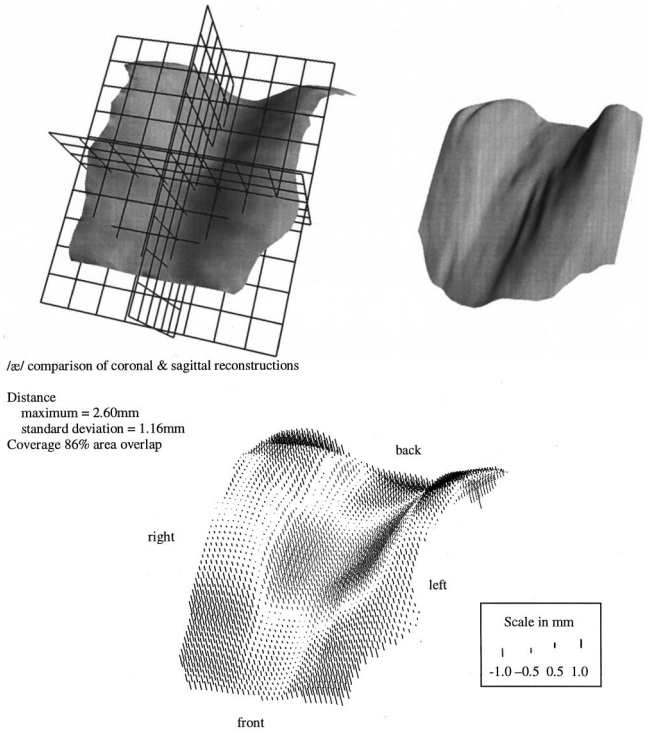

FIG. 1. Reconstructions of coronal /æ/ surface (left) and sagittal /æ/ surface (right) and distance vectors showing the distances between them.

For comparison, and to test measurement repeatability, the same judge twice measured the /ɑ/ surface of a single coronal ultrasound image (see Fig. 2). A year had passed between the two measurements of the images. The error distance between the two reconstructed surfaces had a maximum error of 1.84 mm, standard deviation error of 0.32 mm, and 96% overlap. The maximum error was greater than the error induced by using sparse reconstructions.

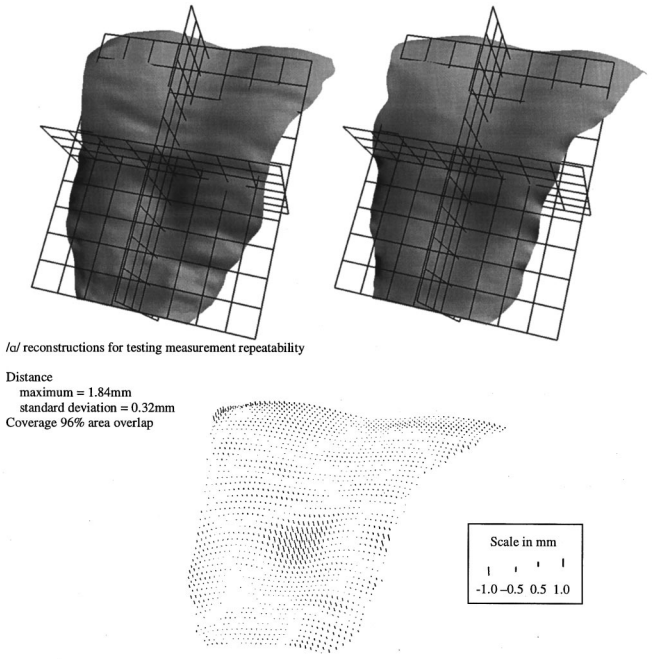

FIG. 2. Two reconstructions of an /ɑ/ measured twice from the coronal images to test measurement repeatability and the distance vector surface between them.

C. Determining an optimal sparse source set

For the dense data sets, tongue surfaces were reconstructed as a b-spline surface that interpolated the dense set of coronal tongue surface contours (Stone and Lundberg, 1996). As tongue surfaces are fairly smooth, particularly between the measured contours of the dense data set, a b-spline surface is a sufficient model. For the sparse data sets, tongue surfaces were similarly reconstructed by defining the b-spline surface that most smoothly interpolated the few coronal tongue surface contours (measured from ultrasound slice images). Tongue surfaces are simple enough to be reconstructed by a small set of coronal contours by this method, but the position of these coronal slices becomes important. For example, for an arched surface the coronal slices must be selected near the point of maximal curvature and displacement. If inappropriate coronal contours are used to reconstruct the tongue, the resulting surface may (at worst) intersect the true surface only along those contours, and may be of a significantly different shape. Selection of a reasonable set of coronal contours is critical to sparse reconstruction of the tongue.

A sparse reconstruction contains just a few coronal slices from the dense set of 60 coronal slices, so there are many possible sparse sets one could collect. In fact, the dense data set can be considered to be 56 slices, as none of the speech sounds had measurable data beyond the 55th slice and slice numbering starts at 0 (see Fig. 3). We considered selecting six slices because this was in fact determined to be the most appropriate choice for balancing data collection constraints and reconstruction accuracy (see Fig. 4). There were then 56 choose 6 [6 selected from 56 without regard to selection order$=56!/(6!(56-6)!)$], or about 32 million possible sets of six coronal slices. The optimal slice set had to be defined globally for all 19 speech sounds, even though each sound had a different optimal set, because the transducer is fixed during actual speech production. There were two desirable properties used in defining an optimal sparse reconstruction. The first was maximal reconstruction coverage, i.e., the ratio of the tongue surface measured in the dense set of tongue slices that was covered by the sparse set. The second was minimizing error.

1. Reconstruction coverage

As the tongue moved forward and back in the mouth during speech, the first and last measurable coronal images (for a fixed dense set of radial images) varied widely (see Fig. 3). Loss of the front slice(s) occurred when the tongue was pulled back and up, creating a sublingual air cavity. In the back, limits on measurable slices were not from tongue position, but from reduced image clarity caused by the increasingly oblique orientation of the tongue surface to the ultrasound beam (which varied in different tongue surfaces and speech sounds). The sound /i/ exemplifies both these problems, as /i/ is the highest of the front raised vowels, and is also very oblique and difficult to measure in the back.

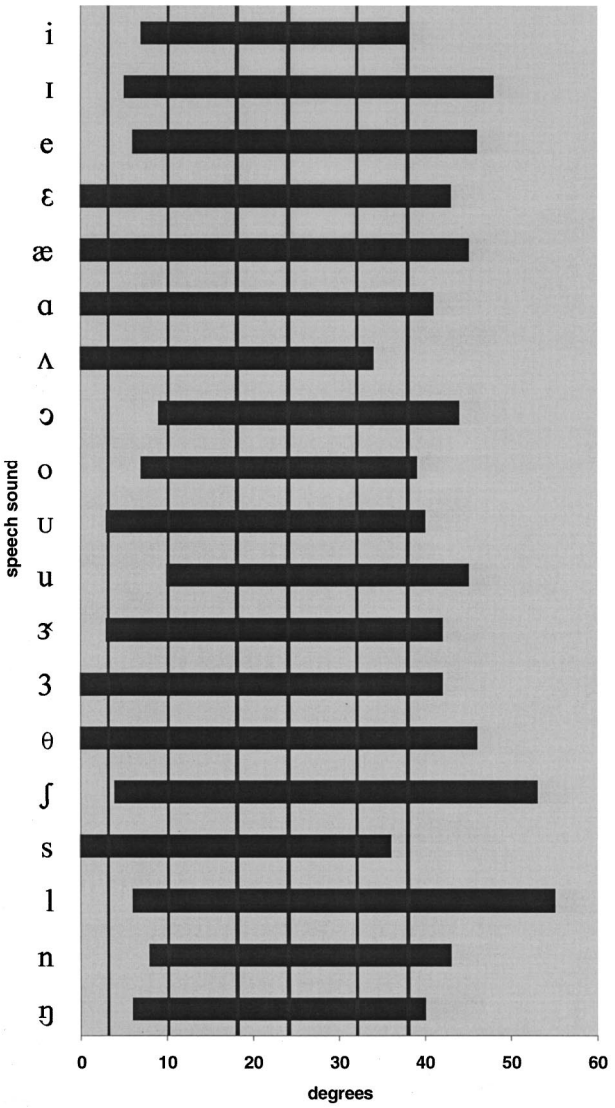

FIG. 3. The range of measurable slices for each of the data sets and vertical lines showing the locations of the six-point optimal sparse source set of coronal slices.

Each of the speech sounds had a specific range of measurable surface contours within the 60-degree 3D sector. For any speech sound, if the extremes of that measurable range are in the sparse data set, the sparse reconstruction of that speech sound will cover the full 3D sector range that a dense data set covered. If the extreme measurable slices are not part of the sparse set, the sparse reconstruction will be truncated at the most extreme slice that does lie within its measurable range. Reconstruction coverage is the area ratio of the sparse reconstruction over the dense reconstruction. As the coronal slices were collected in a polar sweep, the reconstruction coverage can be estimated by the degree range covered by the measurable slices of the sparse set divided by the range covered by all measurable slices for any specific sound. For example, for /i/ only four of the six slices fell within the measurable surface (see Fig. 3), so its reconstruction coverage is 25 degrees/31 degrees.

In order for a 3D reconstruction to be useful it should cover as much of the tongue as possible. Therefore, sparse sets containing from two to nine-slices were optimized for maximum coverage (see Fig. 4). The benefit gained from increasing the number of slices diminished beyond six slices. A second consideration was the practical limitations of real speech data collection. Using current ultrasound instruments, the subject must repeat the speech corpus once for each slice, as they are collected independently. Therefore, fewer slices are preferred. The third consideration was a prior indication that a large reduction in the number of slices was feasible (Stone, 1990). Based on these three considerations and the data in Fig. 4, six-slice sets were optimized.

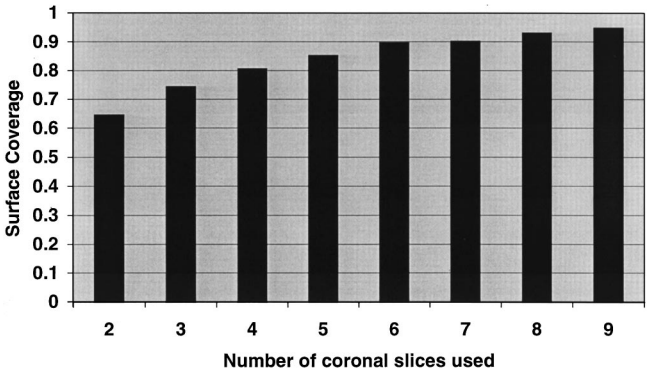

FIG. 4. Coverage of reconstructions for sparse data sets with different numbers of coronal slices.

2. Error analysis of the six-slice set

The second property desired in an optimal reconstruction was minimal error. Sparse sets of six slices were optimized for minimum error. Reconstruction of the tongue surface from a sparse set of slices was identical to the method for reconstructing from a dense data set. An interpolating b-spline surface was fit to the set of surface data points. For the sparse data set from six coronal slices, this had the effect of simplifying the reconstructed surfaces along the sagittal axis. The resulting tongue surfaces were smoother, and might lose detail. To measure the errors induced by this data reduction, the dense reconstruction was compared to the sparse one. To do this, a regularly spaced 2D grid of vertical lines (about one 1-mm spacing) was intersected with the tongue surface. A large enough grid was selected so that all the tongue surfaces in the data set were covered. These intersections gave a regularly spaced set of tongue surface points from the dense reconstructions. For each grid point in the dense data set reconstruction, the closest surface point was found for the sparse reconstruction. The 3D distances between these point sets over all points gave a set of errors. From these, maximum and standard deviation errors measured in millimeters were determined for each of 19 tongue surfaces (corresponding to 19 different static speech sounds). The 3D distances were used as a distance measure, as purely vertical error measures would exaggerate the distances for oblique areas of the surface. In contrast, 3D error distances are measured in a direction normal to the surfaces in all areas.

TABLE I. Reconstruction errors resulting from sparse reconstructions based on different optimizations. No optimization (None) refers to simply taking an equidistant spacing of slices over the full range of the data. Errors reported are for 3D error measured over the surfaces of the 19 speech sounds, except for the six-point set.

Optimization	Selected slices	Average error	Maximum error	Standard deviation error	Coverage	Error cost=max+s.d. +[2×(1-coverage)^3]
None	(0 11 22 33 44 55)	0.37	2.66	0.53	0.79	4.21
Six-slice set	(0 8 16 24 33 38)	0.23	1.42	0.32	0.82	2.63
Six-point set	(3 10 18 24 32 38)	0.21	1.40	0.29	0.80	2.67
Six-point set[a]	(3 10 18 24 32 38)	0.20[a]	0.81[a]	0.27[a]	0.80[a]	2.06[a]

[a]Errors measured for the midsagittal contour only.

3. Optimization of global error cost

To select a set of optimal sparse coronal slices, optimality was defined as minimizing the error cost function:

$$\text{error cost} = \text{maximum error} + \text{s.d. error} + [2 \times (1 - \text{reconstruction coverage})^3]. \quad (1)$$

In addition to this cost function, sparse sets covering an average of less than 0.66 of the 19 speech sound set were eliminated from consideration to prevent the optimization from being skewed by outlying maximum errors. The constant 2 in the error cost equation balances the optimization between the goals of minimizing error and maximizing coverage. Some balance is necessary because simply maximizing coverage results in high surface errors (larger errors than the default equidistant errors in Table I), and optimizing for error only would result in shrinking the sparse surface to consecutive slices one degree apart. For all subjects presented in this paper, the value 2 worked well for both sagittal contour and 3D surface optimizations. Evaluation of the error cost for any slice set required a sparse reconstruction for each of the speech sounds. A brute force search of the 32 million possible sets would require roughly a year of processing time. Thus, a search was needed that could give a fast approximation to the global optimum. A coarse to fine method was used to first get a rough estimation of the global optimum, and then refine that estimation. To do this, the method tests all possible six-slice sets with the restriction that only every fourth possible slice from the dense set is considered. This is equivalent to finding the best sparse reconstruction from a set of coronal slices space 4 degrees apart. So, at this most coarse level, there are only $56/4=14$ possible slices. Since 14 choose 6 is only 3003, all these possibilities can be tested. After determining this coarse step optimum, six-slice sets at a finer level are tested. Now restricted to every second slice, all possible six-slice sets within a single size 2-degree step are considered. In other words, at each slice position, consider the slice 2 degrees before, the current slice, and the slice 2 degrees after, and choose the best permutation across all six slices. Thus, selecting from three slices at six positions gives $3^6=729$ permutations to consider. The best of these permutations is then refined by the same process, using a step size of 1 degree, to give a global optimum approximation. This coarse to fine method is much faster than considering the full range of permutations. It considers $3003+729+729=4461$ rather than 32 million possibilities. The possibilities it does not consider are those where multiple slices are within 4 degrees of one another. For creating sparse data sets, it is highly unlikely that such data sets will be optimal (over the total range of 50+ degrees), so the use of the coarse to fine method should be very reasonable.

The optimal sparse coronal set, for all 19 sounds, resulted in an average error of 0.25 mm, a standard deviation of 0.33 mm, a maximum error of 1.47 mm, and 84% coverage of the dense data sets. Due to the variability in length of tongue surfaces the maximal reconstruction coverage possible for any six-slice set would be 90% (see Fig. 4). As ultrasound has a measurement error around 0.5 mm, the sparse data set was a very good approximation. This indicated that accurate reconstructions could be made from time-varying ultrasound with as few as six slices (at the appropriate positions).

D. Optimizing source sets for individual subjects

These data sets and analyses were based on a single subject, so there is legitimate concern that any subject's optimal sparse source set will vary based on factors of speech production, subject size, or the surrounding vocal tract shape. It would be foolish and impractical to do a dense 3D reconstruction of each subject simply to find the best sparse slices. For this reason, a simpler method for estimating optimal sparse source sets was sought. Instead of measuring the error in the entire 3D surface reconstruction from a sparse coronal slice set, error was measured only in the 2D midsagittal contour reconstruction from a sparse midsagittal point data set. In effect, this would concentrate on the midsagittal slice, and perform the same analysis as finding the best set of slices but in only two dimensions. This was simulated on the dense data set by extracting the midsagittal profiles for the 19 speech sounds, and determining the optimal set of points needed to best reconstruct the global set of profiles. The coronal slices corresponding to the optimal midsagittal points were remarkably close to those selected by the 3D analysis. Using them as the sparse set resulted in slightly reduced surface coverage, but also gave improved error measurements particularly at midline.

II. RESULTS

The goal of this study was to reduce the representation of the tongue surface to a few key slices (i.e., optimal sparse source slices). These slices had to reconstruct 3D tongue surfaces with the highest accuracy possible. If this step was accomplished adequately, the procedure could be developed further to collect time varying data at each slice for use in 4D

TABLE II. Errors in 3D reconstructions based on the optimizations of six coronal slices and six midsaggital points.

	6 slices				6 points			
Sound	Average error	Maximum error	Standard deviation error	Coverage	Average error	Maximum error	Standard deviation error	Coverage
i	0.37	1.42	0.48	0.97	0.32	1.37	0.44	0.90
ɪ	0.17	0.88	0.22	0.70	0.17	1.03	0.23	0.65
e	0.16	0.82	0.21	0.75	0.15	0.51	0.20	0.70
ɛ	0.23	1.34	0.31	0.88	0.26	1.40	0.32	0.81
æ	0.26	0.77	0.28	0.94	0.19	0.72	0.25	0.83
ɑ	0.19	0.79	0.25	0.93	0.21	1.06	0.28	0.85
ʌ	0.42	1.42	0.52	0.97	0.31	1.18	0.40	0.85
ɔ	0.32	1.20	0.41	0.73	0.28	1.29	0.36	0.93
o	0.16	0.74	0.21	0.83	0.18	0.70	0.24	0.73
ʊ	0.20	0.89	0.29	0.81	0.12	0.49	0.15	0.95
u	0.21	0.86	0.27	0.63	0.17	1.14	0.24	0.80
ɝ	0.38	1.38	0.50	0.80	0.20	1.06	0.27	0.90
ɜ	0.14	0.53	0.17	0.91	0.21	0.79	0.28	0.83
θ	0.15	0.83	0.20	0.83	0.15	0.60	0.19	0.76
ʃ	0.32	1.24	0.42	0.73	0.29	0.99	0.36	0.68
s	0.18	0.64	0.22	0.92	0.16	0.52	0.20	0.81
l	0.22	1.07	0.30	0.61	0.24	0.90	0.30	0.57
n	0.20	1.38	0.30	0.86	0.19	1.14	0.27	0.80
ŋ	0.22	0.86	0.29	0.88	0.27	0.84	0.34	0.82
Range	0.14–0.42	0.64–1.42	0.17–0.52	0.61–0.97	0.12–0.31	0.49–1.40	0.12–0.44	0.57–0.95

reconstructions (x,y,z,t). Two sparse source sets were considered. The first was the set of six coronal slices optimized from all the coronal slices of the 56-slice dense set (hereafter called the six-slice set). The second set was the six coronal slices corresponding to the midsagittal points optimized for reconstructing the midsagittal profile (hereafter called the six-point set).

A. Global characteristics of the reconstructions

For each of the sparse sets, global measures of reconstruction accuracy were calculated. Table I shows the optimal six-point and six-slice sets with their global reconstruction errors. Maximum error, standard deviation error, surface coverage, and the resulting cost function were calculated for the entire set of surfaces. The results indicated that the best optimal sparse source was the six-point set. Surface coverage was degraded from the six-slice set optimum and error was improved. Use of midsagittal points as a source set tended to produce better reconstructed surfaces than the coronal set, in many cases, because midsagittal points focused the optimization algorithm on midsagittal features. Thus local depressions, or ''dimples,'' as seen in /l/ and /ɑ/, and steep slopes, as seen in /i/ and /ɝ/, were better captured using the midsagittal source sets. Larger error was seen instead at the surfaces' extreme edges (the least important areas) and also in areas of left-to-right asymmetry, as midsagittal optimization ignores and thus may diminish asymmetries. Errors for individual sounds are shown in Table II. The six-point reconstructions had smaller average errors than the six-slice reconstructions for 11 of the 19 sounds; maximum error was smaller for 13 of the sounds, including all the consonants. Coverage was improved in only four cases. The sparse reconstructions were also more accurate than repeated measurements of a frame (Fig. 2) or comparing sagittal versus coronal dense data sets (Fig. 1). This would indicate that human error in edge detection would be the primary source of error in sparse reconstructions. Concurrent with the present study, a new and automated edge detection system is being developed that should improve measurement reliability.

B. Preservation of local features

In addition to global statistical error measurement, four ''local'' features were considered: left-to-right asymmetry, abrupt changes in slope, local surface depressions, and the constriction location in fricatives. Visual inspection of the dense reconstructions indicate that depressions and abrupt changes in slope were most evident in the midsagittal plane (Stone and Lundberg, 1996; Figs. 4–6). Preservation of these two features in the sparse reconstructions was enhanced by optimizing slice selection in the midsagittal plane. A source set determined by midsagittal points cannot account for left/right differences in shape or motion. In these data sets asymmetry was least well represented. If the selected slices passed through maximally asymmetric regions, the length of the asymmetry would be overestimated. If the slices missed the areas of maximal asymmetry, the degree of asymmetry would be underestimated. The most asymmetrical tongue shape in the data set was /i/ where the maximum error was 1.37 mm.

The second and most easily resolved local feature was the local dimple seen in low back vowels and /l/ (Stone and Lundberg, 1996; Figs. 5 and 6). The use of the five- and six-point sparse sets instead of the coronal sparse set greatly improved resolution of centrally occurring depressions in the 3D surfaces, as they were key features in the midsagittal profile as well. Figure 5 compares the dimple in the dense and sparse surfaces for /l/.

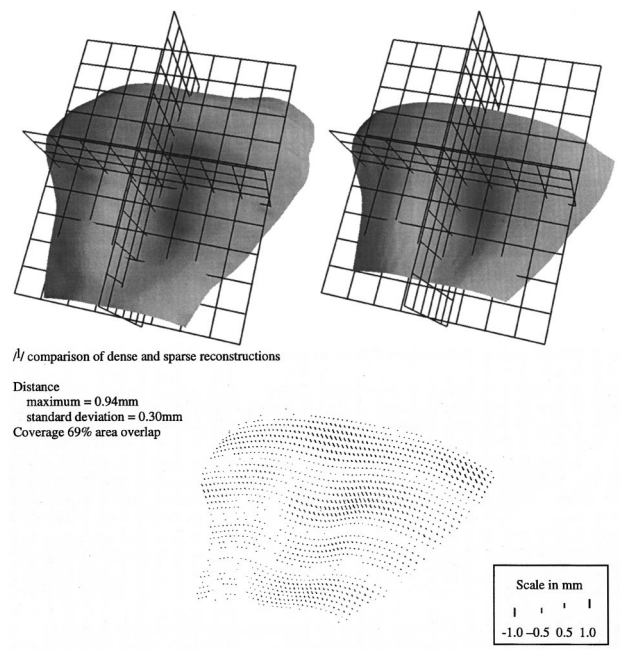

FIG. 5. Reconstructions from a dense set of slices (left), and from a six-slice set (right) for /l/ and the distance vector surface between them.

The third local feature was abrupt change in slope. This feature was particularly evident for /i/ which had an arched tongue in the front, and abruptly became grooved in the back. In addition, the measurable tongue surface was very short. The six-point set resulted in four measurable slices for even the shortest tongue surfaces, and captured the grooves very accurately. Figure 6 shows good representation of abrupt slope changes and deep groove in /i/.

The fourth local feature was the location of fricative constrictions. Fricative constructions in English often occur slightly off midline. Moreover, they may not be marked by

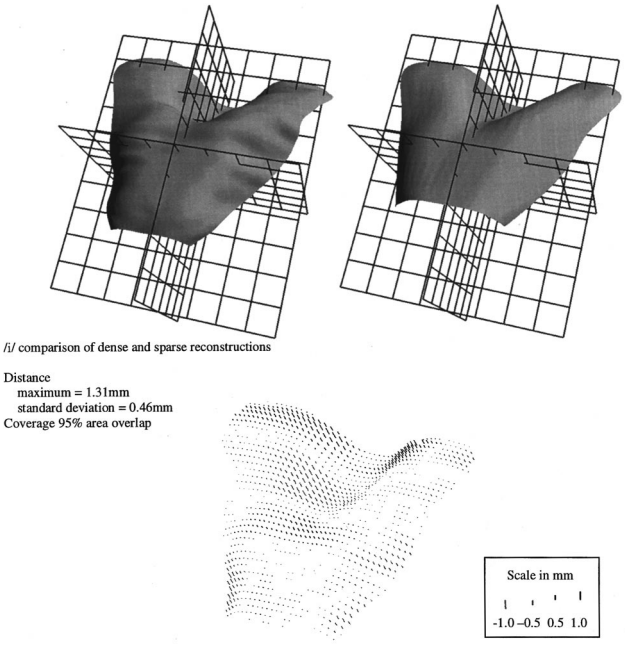

FIG. 6. Reconstructions from a dense set of slices (left), and from a six-slice set (right) for /i/ and the distance vector surface between them.

midline tongue features. Therefore, particular attention was paid to the error in the three fricatives /θ/, /s/, and /ʃ/. Table II shows the maximum error for each sound. For /θ/ the maximum error 0.60 occurred laterally, though not at the edges. Maximum error was 0.2 mm at the constriction. Electropalatography (EPG) data confirmed the subjects constriction locations (see Stone and Lundberg, 1996). For the /s/ the largest error was 0.5 mm and occurred at the edge. At the constriction, the largest error was 0.3 mm. The /θ/ and /s/ shapes were actually fairly easy to predict from a sparse set because the tongue shape did not change dramatically from front to back. The /ʃ/ had a more changeable surface shape and had larger average and maximum errors. The largest error, 0.99 mm, occurred laterally. Several errors of 0.7 mm did appear in the constriction region slightly off the midsagittal plane. In the constriction region, the sparse data set was below the dense data set, which overestimated the channel size.

C. Variability

Intrasubject variation occurs because humans do not say an utterance exactly the same way every time. Phoneme production varies slightly from repetition to repetition. Intrasubject variation could not be seen in these single utterances. One example of variation was contrived, however. The phoneme /l/ was repeated twice by the subject with the goal of creating two different shapes. The first, /l_1/, was produced normally. The second, /l_2/, was produced with a forceful apical contact. Both were sustained about 10 s. The two tongue surfaces were measured and reconstructed using the same procedure as in Figs. 5 and 6. In the /l/ comparison, however, the two surfaces were dense reconstructions of different repetitions (see Fig. 7). We were interested in what causes the midsagittal depression often seen just behind the tongue blade in /l/. It was hypothesized that the more forceful apical contact would create a larger deformation for /l_2/, the more extreme or ''tense'' production. Therefore, the /l_2/ would have a deeper depression than /l_1/. This was found to be true. The lower left portion of Fig. 7 shows the two tongues spatially aligned and superimposed. The /l_2/ tongue is higher than /l_1/ in back and on the sides, and lower in the depression region. The depression depths were 2 mm for /l_1/ and 4 mm for /l_2/, at the deepest point relative to the highest point in the same coronal slice. The important features, the dimples, differed across the repetitions by 2 mm, larger than the maximal error for sparse reconstructions. This number should be accurate since it occurred in the midsagittal region where we generally expect smaller reconstruction errors.

Intersubject variation occurs because humans have slightly different oral morphologies and use different strategies for creating speech gestures. Table III presents midsagittal optimization data from 17 additional subjects. Four speech sounds were collected for each of these subjects: /æ/, /ɑ/, /i/, and /u/. The best six-point sparse set (optimal selections) is compared to the equidistant six-point sparse set (default selections). The optimized selections column presents the optimized range of slices. This makes it clear that across subjects there existed a variety of tongue lengths and feature locations. Smaller ranges were caused by two things: tongues that had incomplete data at one end or the other; and little

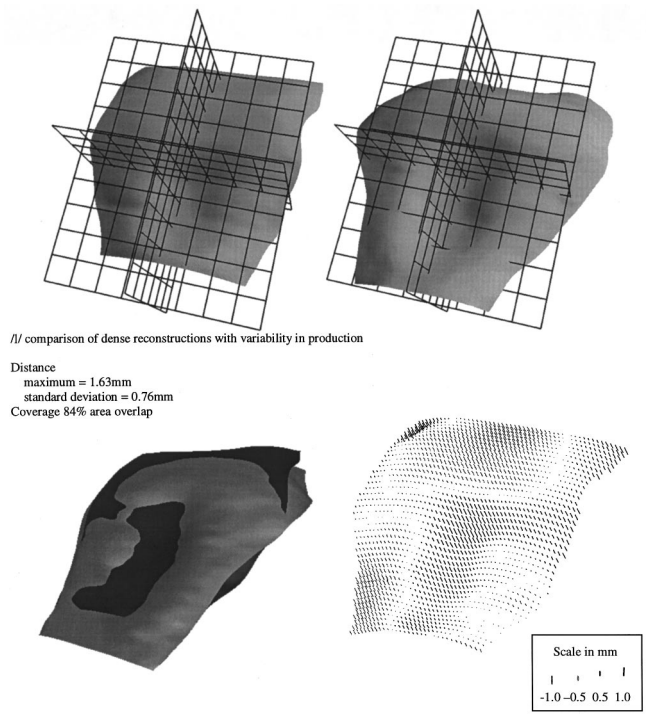

FIG. 7. Reconstructions of dense sets from two distinct productions of /l/ (normal and tense) to show intrasubject variability, the distance vectors between them, and, on the lower left, the two surfaces superimposed in their best alignment. The black surface is for normal production, $/l_1/$, and the gray surface is for tense production, $/l_2/$.

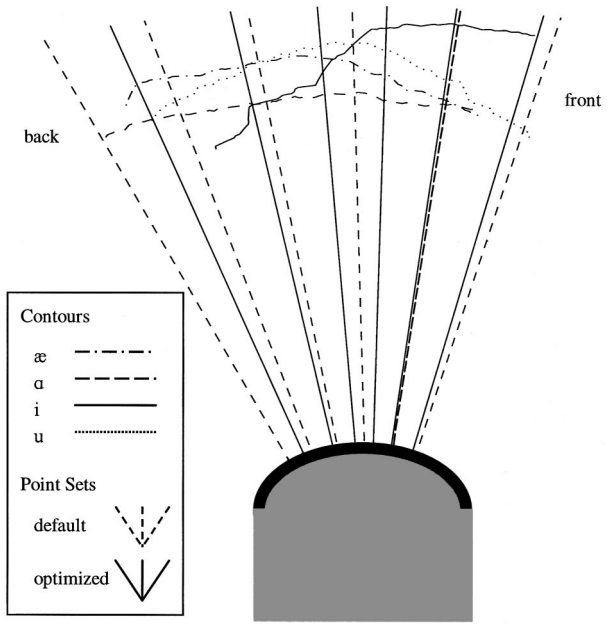

FIG. 8. Sagittal contours of the tongue for subject 7 showing radial lines indicating the default (dashed) and optimal (solid) point sets and their intersections with the four speech sounds.

anterior-to-posterior differences among sounds. Different starting and ending slices among subjects, e.g., subject 3 versus subject 16, reflect rotational differences in positioning the transducer, not true position differences. The primary subject is the subject used in the rest of the paper.

Table III shows that for 11 of the 17 subjects the optimization reduced the maximum error by at least 0.8 mm and for 13 subjects it increased the surface coverage. Subject 7 benefited the most from the optimization. Her default equidistant point set selected points 6 degrees apart with an overall tongue length of 30 degrees. After optimization, her tongue length was 25 degrees, and her interpoint distances were 5, 4, 4, 5, and 8 degrees apart, indicating a greater representation in the anterior tongue. Figure 8 shows the four vowel contours and the optimized points for subject 7 as radial trajectories. With these modifications in point location for this subject, the average error decreased from 0.52 to 0.28, the maximum error decreased from 2.42 to 1.04 mm, and the standard deviation decreased from 0.68 to 0.37. The least improvement was seen for subject 10 whose errors improved only slightly. For some subjects, the optimization was essential, for without it some sounds only had two of the

TABLE III. Midsagittal optimization for the primary and additional subjects based on four speech sounds. The default selections and results (in parentheses) use simple equidistantly spaced points. Values of (---) are displayed for selections that captured only two points for at least one sound, so that no spline estimation can be done for that sound. Errors reported here are measured only on the midsagittal contour.

Subject	Optimal selections	Default selections	Average error	Maximum error	Standard deviation	Coverage
Primary	[00 10 18 23 31 35]	(00 09 18 27 36 45)	0.15 (0.27)	0.45 (1.40)	0.19 (0.38)	0.84 (0.82)
1. A. C.	[10 13 21 29 38 43]	(09 16 24 31 39 46)	0.50 (0.54)	1.49 (1.86)	0.63 (0.70)	0.90 (0.87)
2. C. S.	[12 17 21 29 35 42]	(09 16 22 29 35 42)	0.32 (0.43)	0.94 (2.15)	0.41 (0.56)	0.88 (0.93)
3. E. B.	[04 12 17 21 26 29]	(02 09 16 22 29 36)	0.21 (0.36)	0.71 (1.59)	0.28 (0.48)	0.79 (0.82)
4. E. D.	[08 14 18 23 27 32]	(02 09 16 22 29 36)	0.37 (0.50)	1.16 (2.01)	0.49 (0.65)	0.84 (0.76)
5. E. L.	[05 10 19 27 35 42]	(00 08 17 25 34 42)	0.41 (---)	1.30 (---)	0.52 (---)	0.88 (0.69)
6. E. S.	[07 12 16 21 26 31]	(01 08 14 21 27 34)	0.35 (---)	0.99 (---)	0.44 (---)	0.84 (0.62)
7. F. S.	[10 15 19 23 28 35]	(09 15 21 27 33 39)	0.28 (0.52)	1.04 (2.42)	0.37 (0.68)	0.86 (0.80)
8. J. M.	[06 10 16 23 31 45]	(05 13 21 29 37 45)	0.43 (0.46)	1.40 (2.27)	0.55 (0.65)	0.83 (0.78)
9. J. U.	[03 07 13 19 25 30]	(00 08 16 24 32 40)	0.27 (0.42)	1.02 (2.40)	0.36 (0.57)	0.81 (0.81)
10. K. L.	[02 06 15 22 30 42]	(01 10 18 27 35 44)	0.47 (0.53)	1.65 (1.68)	0.58 (0.68)	0.90 (0.79)
11. K. R.	[16 21 24 28 33 38]	(13 18 23 29 34 39)	0.31 (0.50)	0.99 (2.07)	0.42 (0.69)	0.87 (0.83)
12. M. B.	[17 21 25 29 32 35]	(14 19 24 28 33 38)	0.23 (0.36)	0.71 (1.42)	0.32 (0.51)	0.84 (0.77)
13. R. S.	[15 19 22 24 29 34]	(12 17 23 28 34 39)	0.28 (0.40)	0.98 (1.40)	0.39 (0.56)	0.85 (0.86)
14. S. F.	[07 10 15 17 23 31]	(03 10 17 24 31 38)	0.39 (0.44)	1.44 (1.50)	0.50 (0.53)	0.79 (0.82)
15. S. G.	[01 08 16 24 32 45]	(00 10 20 31 41 51)	0.39 (---)	1.16 (---)	0.48 (---)	0.84 (0.66)
16. T. M.	[19 23 27 32 37 40]	(16 22 27 33 38 44)	0.38 (0.45)	1.33 (1.54)	0.50 (0.63)	0.84 (0.83)
17. V. S.	[18 23 28 31 34 38]	(17 22 26 31 35 40)	0.34 (0.49)	1.18 (1.45)	0.46 (0.62)	0.93 (0.78)

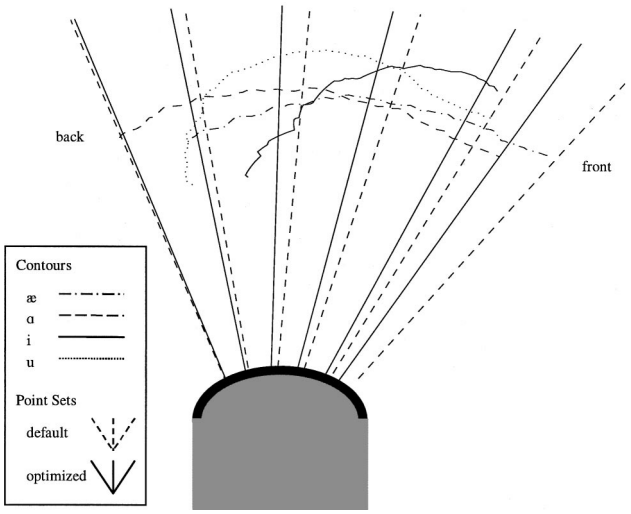

FIG. 9. Sagittal contours of the tongue for subject 5 showing that the default set (dashed) fails to measure the contours for one sound at three (the solid or more points, while the optimal set does intersect each contour at three or more points.

six equidistant points fall on the tongue surface, as marked by (---). This occurred when the subject had great anterior–posterior differences in tongue position across sounds. Figure 9 presents an example of this positional difference for subject 5. The contour for /i/ went from a two-point to a three-point representation after the optimization.

III. DISCUSSION

This study was able to reconstruct 3D tongue surface shapes using as few as six coronal slices. The best slice selection used an optimized set of midsagittal points.

Three important issues are involved in choosing a sparse data set for 3D reconstruction. The first and most important issue is finding the best six-point source set for each subject. Without this, results cannot be generalized across subjects and validity of the method is breached. The optimal sparse source sets determined here will certainly not be optimal for all subjects. Therefore, prior to data collection, a midsagittal data set needs to be collected for each subject. From this data set an optimal source point set is determined for reconstruction of the midsagittal profile. Coronal images would then be collected at each point and reconstructed as described earlier. This procedure can be used to collect time-varying speech samples at each coronal slice angle for use in 3D time–motion reconstructions of the tongue surface during speech. The midsagittal data can be used in the reconstructions as well.

Second, the transducer must be positioned in an accurate and precise manner. A positional error of a few degrees in one slice will reduce significantly the capture of local shape features such as dimples and degree of grooving. Although not addressed in this paper, a 3D automated head and transducer support system (AHATS) based on the currently used 2D head and transducer support (HATS) system (Stone and Davis, 1995) is under construction. This system uses computer (or manually) controled positioning of the transducer at predetermined angles for collection of real-time sagittal and coronal motion. Manual transducer positioning is acceptable if predetermined slice positions are calculated accurately, and precision of transducer placement is assured.

The third issue of importance is reconstruction accuracy of 3D shape and motion. Global reconstruction was optimized by minimizing the maximum and standard deviation errors. As a result, the average errors were below the measurement error for ultrasound. The largest maximum error for all 19 sounds was 1.40 mm, which occurred in /ɛ/ one time on an extreme edge. The greatest standard deviation error was 0.44 mm, which occurred for /i/. Errors above 1.2 mm occurred exclusively at the most lateral edge, and errors above 1.0 tended to occur in the posterior most row.

Optimized reconstructions also need to represent local features well, such as asymmetry, local depressions, and steep slopes. Optimization improved representation of local features compared to equidistant slices. Midsagittal optimization further improved midsagittal features. The first feature, tongue asymmetry, is more prevalent in tongue motion than in static data and so will be even more important for future studies. Left-to-right rotation and a ''leading edge'' are seen fairly often in coronal ultrasound images of speech. These asymmetries do not vary systematically with palatal shape, or handedness (Hamlet, 1987); they are more prevalent in some subjects and some tasks, however. When the slice selection is based on midsagittal points, asymmetries cannot be taken into account, since no lateral information is available. However, leading edges and left-to-right rotations extend across a fairly long region of the lengthwise tongue and, therefore, should be captured by one or more sparse slices. Future research will continue to carefully assess error in representation of asymmetry using the current method.

The second feature, local tongue depressions or dimples, was visible in this data set for nonhigh back vowels and /l/. They have been observed fairly often in other ultrasound data sets (Davis, 1996; Fig. 1) and can be inferred from some point tracking data [Stone, 1990 (Table I)] and MRI data (Kumada *et al.*, 1992; Niitsu *et al.*, 1992) as well. They tend to appear in the ''middle'' segment of the tongue (approximately 2.5–4 cm back from the protruded tip) (cf. Stone, 1990). The present 3D reconstructions captured dimples very accurately because dimples occur at midline and the six-point sparse set optimized their representations.

Accurate representation of steep slopes was the third feature examined in the reconstructions. Front raised sounds (e.g., high front vowels) have a very advanced tongue root. This is due to genioglossus posterior (GGP) contraction, which causes a deep posterior groove defined by a steep slope midsagittally and laterally. Anteriorly, the tongue surface is high and flat, or even arched. Therefore, a sharp inflection point in the midsagittal profile separates the anterior arch from the posterior groove. Choosing a point too far from the inflection point will cause a serious underestimation in the slope magnitude and origin point. Moreover, during changes from front raising to other shape categories, such as back raising, 3D motion reconstructions from inappropriately selected slices will misrepresent and reduce the accuracy of the deformation.

One type of ''error'' is utterance-to-utterance variation,

or free variation. Humans do not produce repeated speech sounds identically. We expect that the induced variation in shape between the /l_1/ and /l_2/ (1.63-mm maximum difference) is the same size or larger than would occur in free variation and is consistent with systematic differences due to morphological constraints. If so, such differences would be larger than the maximum measurement error (1.40 mm) caused by using the sparse data set and should be well represented, especially if the important features occur at midline.

The current sparse set criteria minimize the problem of accurate 3D tongue reconstruction from a sparse slice set, as can be seen from the maximum and standard deviation errors in Table II. The standard deviation errors seen in the data were no worse than typical measurement error. The maximum errors (above 1.3 mm) were seen on the edges. Our expectation is that the selection of fairly equidistant slices, and the optimization across all the lingual sounds in English, will continue to provide as reasonable a 3D coverage as is possible.

ACKNOWLEDGMENT

This work was supported by a research grant from the National Institute of Deafness and Other Communication Disorders (NIH-R01DC01758).

Davis, E., Douglas, A., and Stone, M. (**1996**). ''A Continuum Mechanics Representation of Tongue Motion in Speech,'' Proceedings of the 4th International Conference on Spoken Language Processing, Vol. 2, pp. 788–792.

Hamlet, S. (**1987**). ''Handedness and articulatory asymmetries on /s/ and /l/,'' J. Phonetics **15**, 191–195.

Keller, E., and Ostry, D. (**1983**). ''Computerized measurement of tongue dorsum movements with pulsed echo ultrasound,'' J. Acoust. Soc. Am. **73**, 1309–1315.

Kier, W. M., and Smith, K. (**1985**). ''Tongues, tentacles and trunks: the biomechanics of movement in muscular-hydrostats,'' Zoo. J. Linnean Soc. **83**, 307–324.

Kumada, M., Niitsu, M., Niimi, S., and Hirose, H. (**1992**). ''A Study on the Inner Structure of the Tongue in the Production of the 5 Japanese Vowels by Tagging Snapshot MRI,'' Research Institute of Logopedics and Phoniatrics, University of Tokyo, Annual Bulletin, Vol. 26, pp. 1–13.

Miyawaki, K., Hirose, H., Ushijima, T., and Sawashima, M. (**1975**). ''A preliminary report on the electromyographic study of the activity of lingual muscles,'' Research Institute of Logopedics and Phoniatrics, University of Tokyo, Annual Bulletin, Vol. 9, pp. 91–106.

Niitsu, Kumada, M., Niimi, S., and Itai, Y. (**1992**). ''Tongue Movement during Phonation: A Rapid Quantitative Visualization Using Tagging Snapshot MRI Imaging,'' Research Institute of Logopedics and Phoniatrics, University of Tokyo, Annual Bulletin, Vol. 26, pp. 149–156.

Sonies, B., Shawker, T., Hall, T., Gerber, L., and Leighton, S. (**1981**). ''Ultrasonic Visualization of Tongue Motion During Speech,'' J. Acoust. Soc. Am. **70**, 683–686.

Stone, M. (**1990**). ''A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data,'' J. Acoust. Soc. Am. **87**, 2207–2217.

Stone, M. (**1995**). ''How the tongue takes advantage of the palate during speech,'' in *Producing Speech: Contemporary Issues: A Festschrift for Katherine Safford Harris*, edited by F. Bell-Berti and L. J. Raphael (American Institute of Physics, New York), Chap. 10, pp. 143–153.

Stone, M., and Davis, E. P. (**1995**). ''A head and transducer support system for making ultrasound images of tongue/jaw movement,'' J. Acoust. Soc. Am. **98**, 3107–3112.

Stone, M., and Lele, S. (**1992**). ''Representing the Tongue Surface With Curve Fits,'' Proceedings of the International Conference on Spoken Language Processing, Vol. 2, pp. 875–878.

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

Stone, M., and Yeni-Komshian, G. (**1991**). ''Some effects of pharyngealization on tongue shape as seen on ultrasound,'' J. Acoust. Soc. Am. **89**, S1979 (A).

Stone, M., Sonies, B., Shawker, T., Weiss, G., and Nadel, L. (**1983**). ''Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system,'' J. Acoust. Soc. Am. **11**, 207–218.