RECONSTRUCTING THE TONGUE SURFACE FROM SIX CROSS-SECTIONAL CONTOURS: ULTRASOUND DATA

Andrew J. Lundberg[1] *and Maureen Stone*[2]

[1]Department of Computer Science, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, [2]Division of Otolaryngology, University of Maryland Medical School, 16 S. Eutaw Street, Suite 500, Baltimore, MD 21201

ABSTRACT

This work presents a method for reconstructing 3D tongue surfaces during speech from ultrasound data. The method reduces the dimensionality of the tongue surface 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. Earlier work (Stone & Lundberg, JASA 99, 3728-3737, 1996) produced 3D reconstructions of static tongue surfaces from dense sets (60 slices) of 2D coronal tongue contours. Sparse data sets (6 slices) from within the original dense set were used to reconstruct 3D tongue surfaces, which were compared to the 60 slice surfaces. The 6 slice sets of coronal images were determined from an optimized set of midsagittal points. The reconstruction procedure was done in an identical manner to the dense data, but in 2D. Cross-sectional slices of the tongue were measured at the "optimal" midsagittal points, and used to reconstruct 3D surfaces. These surfaces were compared to the dense reconstructions. Errors and reconstruction coverage were comparable to the 3D optimized sparse set, indicating this was an adequate method for calculating a sparse data set for use in reconstructing 3D surface behavior.

1. INTRODUCTION

The present study determined a minimal number, or sparse set, of coronal tongue slices needed to reconstruct 3D tongue surfaces (x,y,z) for 19 static English sounds as a precursor to reconstruction of 4D tongue surfaces (x,y,z,t). The goal of this study was to specify optimal sparse sets (OSS) of 6 coronal ultrasound slices to recreate the 3D tongue surfaces reconstructed from dense sets of 55 slices. The optimization procedures discussed below were done initially (not reported here) using the entire 3D data set of coronal slices. Results for that data set can be seen in the last row of Table 1. However, it would be foolish and impractical to collect 55 coronal slices on each subject simply to find the 6 best slices. For this paper, therefore, we present a variant of the original method which uses midsagittal data to estimate the best coronal slices for the sparse reconstructions. There are two advantages to determining the OSS from midsagittal contours. First, every subject's OSS will be different based on factors such as subject size, or the surrounding vocal tract shape. By collecting midsagittal data first and running the optimization program, one can determine and collect the optimal coronal slices during the same recording session. The second advantage is that by concentrating on the midsagittal contour, midline features (such as local depressions and sharp changes in slope) are well captured in the chosen coronal slices.

There is no question that tongue behavior in the midsagittal plane cannot be extrapolated simply into 3D surface behavior. We have seen motion patterns in coronal slices that would not be extractable from midsagittal data (cf. Stone, 1990). However, these 3D reconstructions came from 6 coronal slices, which covered about 83% of the original area of the tongue. The location of the slices was determined from the optimized midsagittal points. The validity of that process is being presented here. We have chosen an extensive corpus of speech sounds, conservative error criteria, and global optimization, to maximize the accuracy of our reconstructions.

2. METHODS

2.1. Subjects and Speech Materials.

The subject was a 26 yo WF with a Baltimore, MD accent. Complete recording procedures and subject information can be found in S&L, 1996. All the sounds of English that use the tongue (or their cognates) were used: /i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɝ, ʒ, s, θ, ʃ, l, n, ŋ/. The dense data sets consisted of from 38 to 55 coronal slices each one degree apart (Figure 1). The OSS sets are represented by the 6 vertical lines in Figure 1.

2.2. Determining an OSS set of Coronal Slices from Midsagitatl Slices:

A sparse reconstruction contains just a few coronal slices from the dense set of 55 coronal slices, so there are many possible sparse sets one could collect. We considered six slices because this was in fact determined to be the most appropriate choice for balancing data collection constraints and reconstruction accuracy. 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 percentage of the tongue surface measured in the dense set of tongue slices that was covered by the sparse set. The second was minimal error.

A mentioned above, midsagittal data was used to determine the OSS sets. 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 midsagittal profiles. The coronal slices corresponding to the optimal sagittal points were remarkably close to those selected by the 3D analysis, and using those slices as the sparse set resulted in minimal global data degradation, and improved midline representation (Table 1).

Reconstruction	Source	Slices	Worst	Average	%
Midsagittal Contour Reconstructions	6 Point Sparse Set	3,11,20,28,36,42	1.20	0.34	82.2
Midsagittal Contour Reconstructions	5 Point Sparse Set	3,11,20,28,36	1.20	0.36	74.4
3D Surface Reconstructions	6 Point Sparse Set	3,11,20,28,36,42	1.85	0.36	82.2
3D Surface Reconstructions	5 Point Sparse Set	3,11,20,28,36	1.85	0.38	74.4
3D Surface Reconstructions	Coronal Slice Set	4,10,21,28,36,42	1.80	0.39	83.2

Table 1: Five and six slice sets used in 3D reconstructions with worst errors, average errors, and percent of reconstruction coverage.

2.3. Percentage of Reconstruction Coverage

As the tongue moved forward and back in the mouth during speech, the first and last measurable coronal images varied widely, see Figure 1. For any speech sound, if the extremes of its measurable range were in the sparse data set, the sparse reconstruction would cover the same range as the dense data set. If not, the sparse reconstruction would be truncated at the most extreme slice that did lie within its measurable range (see Figure 1). Reconstruction coverage was the sparse surface reconstruction area in degrees divided by the measurable dense surface area in degrees.

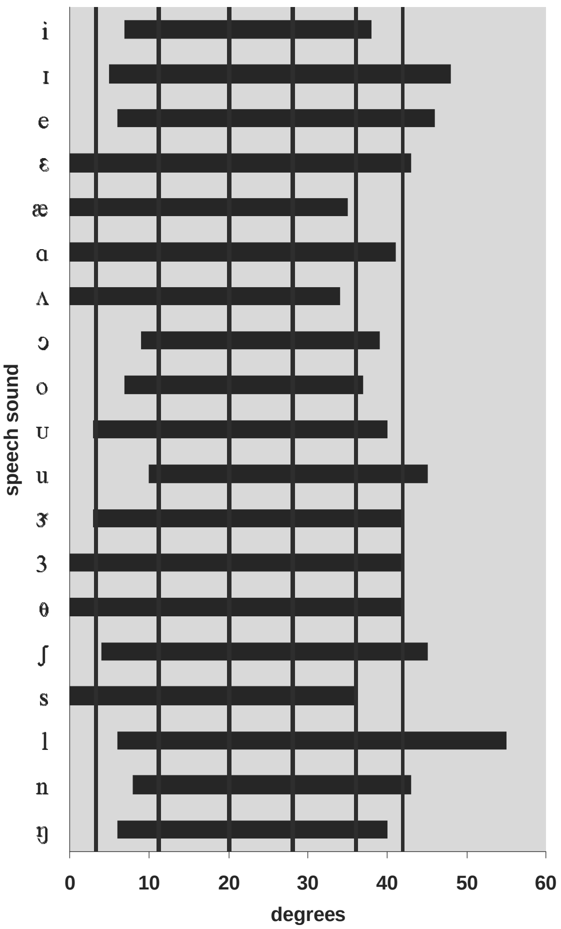

Figure 1: The range of measurable slices for each of the data sets and vertical lines showing the location of the optimal coronal slices

Reconstruction coverage had to be balanced against data collection considerations. Since the subject must repeat the speech corpus once for each coronal slice, the fewest number of slices was preferred. Sparse sets containing from 3 to 10 slices were analyzed for maximum possible percent coverage. The benefit gained from increasing the number of slices diminished beyond six slices and so 6-slice sets were optimized. In addition reconstructions made from 6-*point* sets (i.e. slices whose location was determined by optimizing midsagittal points) and 5-*point* sets (the 5 anteriormost slices) were evaluated.

.

2.4. Error Analysis of the Six Slice Set

The error cost function was a measurement of both the worst and average errors. An interpolating b-spline surface was fit to the sparse set of surface data points. The resulting tongue surfaces were smoother than the dense data set, and might lose detail (Figure 2). The dense reconstruction was compared to the sparse one using a 2D grid of vertical lines that intersected with the tongue surface. 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 gave a set of error responses. From these, worst and average errors were determined for each of 19 speech sounds.

We defined optimal error as minimizing an error term composed of worst error, average error, and percent coverage. This error term was considered over the reconstructions of all 19 speech sounds, defining a global optimum. To quickly find a good global optimum, a technique called simulated annealing [5] (Kirkpatric et al., 1983) was used. The simulated annealing process seeks a global optimization by allowing random movements in the search space of all possible 6 slice sets.

3. RESULTS

The goal of this study was to reduce the representation of the tongue surface to a few key slices . This procedure can be developed further to collect time-varying data at each slice for use in 4D reconstructions (x,y,z,t).

Error results for reconstruction of the midsagittal profiles from 5 or 6 points, and the reconstruction of the 3D surfaces from 5 or 6 slices appear in Table 1. 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).

3.1. Global Characteristics of the Reconstructions

For each of the sparse sets, global measures of reconstruction accuracy were calculated. Table 1 shows the OSS data derived from 5-*point*, 6- *point*, and 6-*slice* source sets, with their global reconstruction errors. Errors, surface coverage, and cost functions were calculated for the entire set of surfaces. The results indicated that the best OSS was the 6-*point* set. Only percent coverage was degraded from the 6-*slice* set optimum. The 5-*point* set introduced a further reduction in percent coverage due to the loss of posterior surface area. In fact, 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. Increased 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).

3.2. Preservation of Local Features

In addition to global statistical error measurement, preservation of important physical features was given weight in determining the error cost. The sparse source set was optimized to minimize local and global 3D reconstruction error. The error cost used in the minimization was selected to preserve as many local features as possible. The three "local" features considered were left-to-right asymmetry, abrupt changes in slope, and local depressions or “dimples.”

Left-to-right asymmetry was diminished in the sparse reconstruction when the selected slices were not in maximally asymmetric regions. Of the three features, asymmetry was the least resolvable. A source set determined by midsagittal points cannot account for left/right differences in shape or motion. Figure 2 exemplifies this loss.

The second and most easily resolved local feature was the local dimple seen in low back vowels and /l/ (Stone & Lundberg, 1996, Figures 5, 6). The use of the 5- and 6-point source sets instead of the 6-slice source set greatly improved resolution of centrally occurring depressions in the 3D surfaces as they were key features in the midsagittal profile as well.

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 sparse tongue surface was very short. Thus, initial global optimization using 5 coronal slices selected only 3 within the measurable range for /i/. As a result, the surface was smoothed excessively and both maximal curvature and slope steepness were reduced. The use of the 5- and 6-point source set resulted in four slices for even the shortest tongue surfaces, and captured the grooves very accurately.

4. DISCUSSION

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

Two important issues are involved in choosing a sparse data set for 3D reconstruction. The first issue is reconstruction accuracy of the 3D surface and eventually motion. Global reconstruction was optimized by minimizing the average and the maximum error. The largest maximum error for all 19 sounds was 1.85 mm. The second issue is finding the best 6-point source set *for each subject*. Choosing coronal slices by optimizing midsagittal points acts to normalize the technique for each subject. Without this, results cannot be generalized across subjects and validity of the method is breached. Coronal images are collected at the subject's optimized point locations and reconstructed as described earlier. The midsagittal data can be used in the reconstructions as well.

Local reconstruction features such as asymmetry, local depressions, and steep slopes were considered in the error analyses, because these are among the most important features of surface shape. 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. When the slice selection is based on midsagittal points, asymmetries can not be taken into account, since no lateral information is available. However, left-to-right asymmetries extend across a fairly long region of the lengthwise tongue, and therefore, should be captured by one or more coronal slices the worst symmetry error in these data appears in Figure 2.

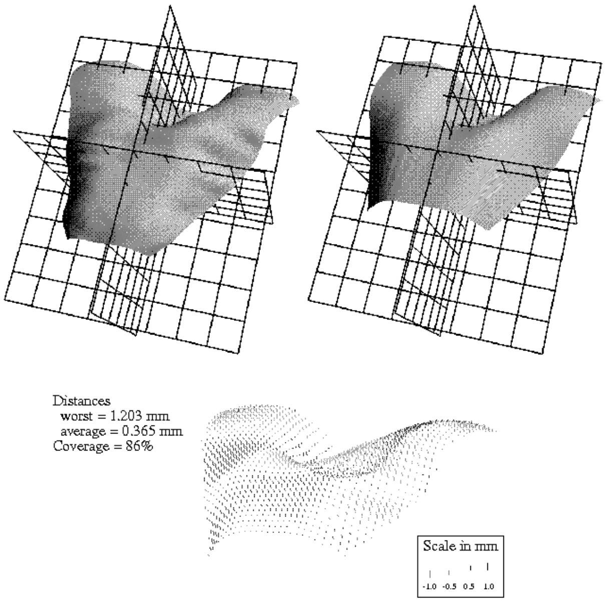

Figure 2: Reconstructions from a dense set of slices (left), and from a 5 slice set (right) for /i/ and the distance vectors between them.

The second feature, dimples, were visible in this data set for non-high back vowels and /l/. The present 3D reconstructions captured dimples very accurately because dimples occur at midline and the 5- and 6-point source sets maximized their representations.

Accurate representation of steep slopes was the third feature examined in the reconstructions. Front raised sounds (e.g., high front vowels) tended to be short in length with a deep posterior groove defined by a steep slope midsagittally and laterally. Anteriorly, the tongue surface was high and flat, or even arched. Therefore, a sharp change in slope in the midsagittal profile separated the anterior arch from the posterior groove. Choosing a slice too far from the change caused a serious underestimation in the slope magnitude and origin point. This problem was resolved adequately by using 6 slices, which allowed a shorter distance between slices, and included at least 4 slices within the body of the tongue for the short, front-raised vowels.

5. SUMMARY

Three-dimensional tongue surfaces were reconstructed accurately from 5-6 coronal slices that were selected using an optimized set of midsagittal points. Overall errors were minimal and located in regions of lesser importance, such as the lateral edges. Local features of importance such as depressions, asymmetries, and changes in slope were captured well. Reconstruction accuracy was within 2 mm maximum error. These sparse representations successfully reduced the dimensionality of the tongue while retaining essential local features. Moreover, these representations did not impose an a priori model on the tongue's shape or motion. In future studies, time-varying 2D ultrasound data collected at these six slice locations are expected to provide adequate reconstructions of 3D tongue surface motion (4D).

6. REFERENCES

Kirkpatrick, S., Gelatt, Jr. C. D., and Vecchi, M. P., "Optimization by simulated annealing," Science, 200(4598), 671-680 (1983).

Stone, M., "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," Journal of the Acoustical Society of America 87, 2207-2217 (1990).

Stone, M., and Lundberg, A., "Three-dimensional tongue surface shapes of English consonants and vowels," Journal of the Acoustical Society of America 99(6), 3728-3737 (1996).