Beyond 2D in articulatory data acquisition and analysis

Philip Hoole[1], Andreas Zierdt[1] and Christian Geng[2]

[1]Institut für Phonetik und Sprachliche Kommunikation, Munich

[2]Zentrum für Allgemeine Sprachwissenschaft, Berlin

Email: hoole|andi@phonetik.uni-muenchen.de, geng@zas.gwz-berlin.de

ABSTRACT

The potential of the extension of traditional EMMA recordings to 5D in a new articulographic system is illustrated. "5D" means that each sensor delivers three translational and two rotational coordinates. This high information density gives the following advantages: lateral tongue movements are a source of information rather than error; subject's head does not need to be fixed relative to the apparatus; the six degrees of freedom of rigid structures can be efficiently captured; rotational information provides additional constraints on the possible shape of deformable structures such as the tongue. Regarding analysis, we discuss the use of 3-way statistical models in recent developments of the PARAFAC tradition, specifically considering the application to multispeaker 3D tongue surface data from MRI scans, an area that presents an interesting and largely unexplored challenge to these models. In addition, a new algorithm for deriving tongue surfaces through efficient merging of axial and coronal scans is presented.

1. INTRODUCTION

This contribution consists of two parts. In the first part, we illustrate the potential benefits for capturing articulatory data offered by the new 5D articulograph currently under development in collaboration with Carstens Medizinelektronik. In addition, we give a first assessment of its performance under realistic conditions by comparing data from the traditional 2D with the newer system, recorded on the same subject without removing the sensors between the 2D and 5D blocks of trials. In the second part of the paper we discuss approaches for parameterizing the 3D surface of the tongue. In particular, we present a new algorithm for merging information from coronal and axial MRI scans to provide a comprehensive representation of the tongue surface, and also discuss possibilities for analyzing this class of data with N-way statistical techniques.

2. FIVE-DIMENSIONAL ARTICULOGRAPHY

A basic outline of the system design has been given elsewhere [8]. Bench-tests of accuracy will also be the subject of a separate report (at the time of writing, accuracy is on the order of 1mm but not completely stable over all possible orientations, so still with room for improvement). Here, we will concentrate firstly on illustrating specifically phonetically useful features of the system, emphasising in particular information gained from sensor orientation, and secondly on a comparison of data acquired in the same session with the 2D EMMA. One of the problems at the start of the development of the latter system was that it was hardly possible to check reliability under realistic conditions through parallel measurements with some independent system. For the new system we now at least have the advantage of a close comparison with the older system, since the systems use the same sensors.

2.1 SENSOR ORIENTATION AS ADDED VALUE

In the new system each sensor provides five coordinates: x/y/z positions and two rotations (azimuth and elevation). This leaves one rotational degree of freedom unaccounted for.

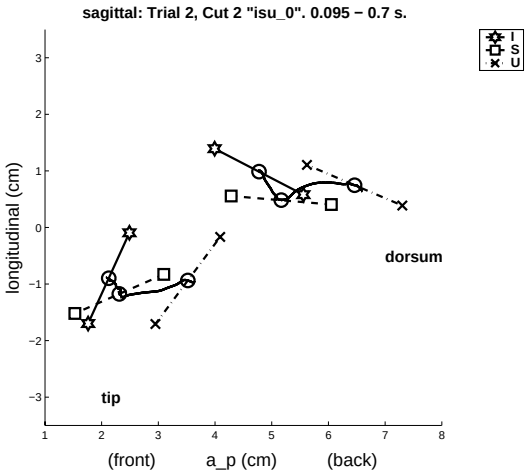

Figure 1: Trajectories of tip and dorsum for /isu/. Bars with symbols indicate sensor orientation at selected time instants.

The example in Fig. 1 illustrates the use of the new orientational information. It shows trajectories of tongue-tip and dorsum in a traditional sagittal view. In this recording the sensors were mounted on the tongue with the main axis of the sensor running along the midline of the tongue. Thus measured sensor orientation responds to changes in curvature of the tongue in the midsagittal plane. At the midpoints of the three sounds in this simple VCV sequence /isu/ the position and orientation of the sensors has been marked by bars (position is marked by circle; the symbol at

the ends of the bars codes the target sound).

The point to note is that there is very little movement of the sensors (especially the tip sensor) from V1 to the consonant (i.e the circles at the midpoint of the bars are close together). However, there is a substantial change in the *orientation* of the sensor, consistent with the change from a bunched tongue configuration with lowered tongue tip for /i/ to raised tongue-tip for the consonant (in fact, the tip of the tongue is actually about 2cm anterior to the 'tip'-sensor).

With only the positional information from the two tongue sensors it would be impossible to realistically assess the configuration of the tongue. With the orientation information we get about as much information on the configuration of the tongue from only two sensors in the new system as we would have got from four sensors in the old 2D system. Conversely, with four sensors in the new system we would expect to obtain a more reliable and detailed picture of tongue shape than before. In particular, it is worth noting that it is not possible to locate a sensor right on the tongue tip (because of disturbance of articulation) so the orientation information will help to give information about locations on the tongue that cannot be monitored directly, and thus should be particularly useful for more precise characterization of laminality vs. apicality. Somewhat similar benefits can also be expected in the tongue root region.

2.2 EFFICIENT EXTRACTION OF RIGID-BODY PARAMETERS

Orientation information is equally useful for rigid bodies. In this section we consider the estimation of the three translational and three rotational parameters characterizing rigid body motion. Motion of the head, for example, is communicatively interesting in its own right, but also needs to be factored out of articulator movement. This is particularly crucial in the new system in order to derive maximum benfit from the fact that the subject no longer needs to be attached to the transmitter assembly. In purely positional measurement systems at least three non-coplanar points are required to solve for the 6 degrees of freedom. Algorithms for use with 3D point data are readily available (we use that in [2]). In order to be able to use such procedures here we split up the 5D information into the 3D spatial coordinates of the sensor itself plus the 3D spatial coordinates of a "virtual sensor" located some fixed distance (e.g 4cm) from the actual sensor along the line defined by the two rotational coordinates (in effect the position of the bar-ends in the previous figure). This in turn made it feasible to solve for the full 6 DOFs of the head using just the two reference sensors used as standard to compensate for head movement in the 2D system, i.e located on the upper incisors and the bridge of the nose. Precisely this procedure was followed in the above-mentioned recording in which the same subject performed the same movement tasks first with the 2D system and then, about 30mins. later, with 5D system. Although the subject did not move her head a great deal, nonetheless over about 150 trials translational movements of about 6cm and rotational movements of about 10 deg. occurred, which would be more than enough to wreck any articulatory analysis if not accounted for correctly. Impressionistically, the outcome was quite successful. The next figure shows 10 overlaid trials for a lower-lip movement task in two bite-block conditions performed with both systems. There is no evidence of the 5D system showing a higher level of unsystematic variability. Further inspection of these figures reveals additional interesting similarities and differences. The absolute spatial position of the sensors does not appear to be completely identical in the two systems. This could reflect misalignment problems in the 2D system or weaknesses in the calibration of either system that cannot be resolved here. On the other hand, there are striking similarities in subtle details of the relative pattern of articulator movement: both systems show a lower position of the upper-lip in the large bite-block condition, compared to the clenched-teeth condition (together with a lower maximum elevation of the lower-lip), indicating attainment of lip closure at a lower vertical position. In labiodental movement tasks (not shown here) the systems agreed in showing an absence of these lip-position shifts between the two conditions, which phonetically makes perfect sense.

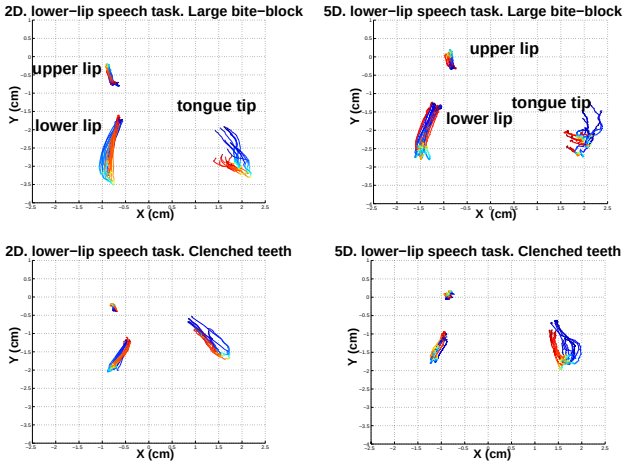

Figure 2: Comparison of articulator trajectories from 2D system (left) and 5D system (right) during articulation of /pa:p/in bite-block condition (top) and clench-teeth condition (bottom). Same axes scaling in all panels.

A more formal way of estimating the success of this registration procedure is to measure the distance at each time instant between the transformed position of the reference sensors and their positions for the chosen reference configuration of the head. This gives a measure of the geometric distortion present in the system. In the present case, this distance measure (averaged over the two true and two virtual sensors at each time instant) gave an average of about 0.4 mm over 150 trials (with a worst case of about 0.7mm). This level of relative accuracy is quite encouraging, but it was nevertheless apparent that the higher levels of distortion occurred systematically as the head moved further away from its reference position. This reflects the fact that the last word on the calibration has certainly not been said. However, the approach outlined here also offers great promise for the other main rigid body in speech, namely the jaw. Resolving jaw motion into translational and rotational

components has been a persistently awkward problem in 2D measurement systems.

2.3 CAPTURING UNEXPECTED LATERAL MOVEMENT

We suspect that the bulk of recordings done with the new system will continue to be based on the midsagittal plane. Nevertheless, it is important that the system catches lateral movement when it occurs. We give here two brief examples where lateral movement was not necessarily expected, and where it could lead to measurement difficulties in the 2D system. In one recording session, for the sequence /alu/ it was observed that as the tongue raised from /a/ to /l/ it also moved laterally about 7mm and changed orientation by about 30 degrees. The /l/ here was most likely an apical articulation; many experienced EMMA users appear to have encountered problems with possibly unreliable data (as indicated by the misalignment factor) with respect to apicals.

A second example of unexpected lateral movement occurred in the experiment that provided the data for the 2D/5D comparison: It was observed that in the bite-block condition (the bite-block was inserted unilaterally) the tongue-tip and lower-lip were displaced laterally by almost 1cm.

2.4 FURTHER 2D/5D COMPARISON: SENSOR ALIGNMENT

Fig. 2 above allowed a comparison of positional information from the 2D and 5D system. It is also possible to compare them to a certain extent with respect to sensor alignment. Of course, the whole point of the 5D development is that the traditional 2D system does not provide specific information on sensor orientation. However, the old system does provide an estimate of the amount by which the sensors are misaligned with respect to the transmitters (in the 2D system ideally they should be parallel), without however indicating how this misalignment comes about. Nevertheless, if the 5D system is behaving properly it should at least be able to confirm the misalignment estimates of the 2D system. This turned out to be indeed the case. Both systems indicated exactly the same rank order of the 5 sensors in use in the experiment: most misalignment (averaged over more than 100 trials) for tongue-tip, least for upper-lip and nose. Moreover the systems both indicated more *change* in alignment for the tongue-tip than the lower-lip during movement tasks, and the same systematic relationship between amount of misalignment and the movement: tongue-tip showed increasing mislignment during opening gestures whereas lower-lip showed decreasing misalignment.

3. STATISTICAL MODELLING OF 3D TONGUE SHAPE

3.1 N-WAY STATISTICAL ANALYSIS

Work in the PARAFAC tradition initiated by Harshman [3] has proved very revealing for articulatory analysis, especially of tongue configuration in vowels. PARAFAC is a 3-mode technique that resolves the rotational indeterminacy of standard Principal Component analysis, combining parsimony for multi-speaker data with a high degree of interpretability. These desirable features may not be achievable in practice because the model places very strong constraints on speaker-specific characteristics. Because of both past experience and anticipated difficulties with multi-speaker MRI datasets (see discussion in [4]) we started to explore further models [5]. PARAFAC is actually just one member of a whole family of 3-way or indeed N-way techniques. Comparison with the so-called Tucker3 model is particularly instructive ([6], [5], [7]). PARAFAC can be seen as one very highly constrained case of this model; thus there exist numerous possibilities for relaxing the constraints. Interestingly, however, Zheng et al. [7] very recently succeeded in fitting a 2-factor PARAFAC model to MRI data of American vowels, so this kind of data may not be as intractable as we originally thought. Nonetheless, their tongue reconstructions were limited to upper tongue surface from coronal scans, so there is certainly still scope for exploring how far one can push highly-constrained models.

3.2 3D SHAPE FROM MULTIPLE MRI VOLUMES

Currently, it is difficult to generate satisfactory representations of the complete tongue shape from MRI volumes taken in only one orientation. (See Badin et al., 2002, for one elegant way round this problem).

Our approach is based on complete coverage of the whole vocal tract in all three traditional volume orientations, coronal, axial and sagittal. Acquistion details are given in [4]. Modelling efforts are currently focusing on vowels (though some consonants have also been recorded). We here outline a procedure for generating 3-dimensional tongue surfaces suitable for input to statistical procedures based on a spherical representation of the tongue (i.e an extension of the old midsagittal tradition of regarding the tongue body as a circle). This procedure is coupled with a technique for merging data from the different volume orientations.

Intial steps consist in determining the alignment of the midsagittal plane based on anatomical landmarks, and in mapping all vowels to a common jaw position. Next, the controid of the tongue mass on the midsagittal plane is used as the origin of a set of spherical coordinates, this origin being determined separately for each vowel. Essentially, a semicircular grid defined in the coronal plane is rotated to a set of positions covering the full vocal tract about an axis perpendicular to the sagittal plane. This results in a rectangular grid (currently 57*114 points), each point on the grid being specified for azimuth and elevation. A subsampled version of the left half of the grid is shown in Fig. 3.

The most arduous task was to define the tongue contour in every individual slice. Basically the procedure involves setting sufficient control points (either manually in difficult regions, or semi-automatically in regions with a clear tissue-air interface) to allow spline reconstruction of a complete contour with points at equidistant intervals of 2mm.

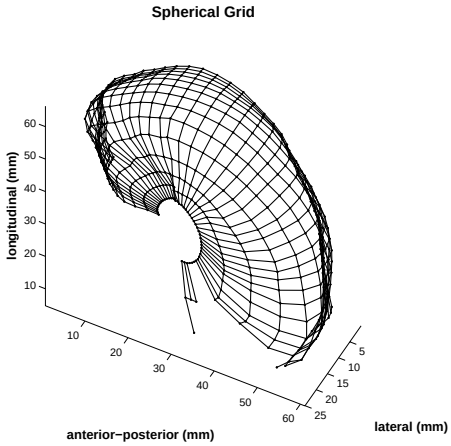

Figure 3: Spherical grid applied to tongue surface of vowel /ø:/ (left half only; some missing data in pharyngeal region)

This raw contour data is then also converted from Cartesian to spherical coordinates relative to the tongue-centre position. It is then possible to use standard surface fitting routines to estimate the radial coordinate of the tongue surface at each of the points on the spherical grid. This procedure is performed separately for data from coronal and axial slice orientations of each vowel. The surface-fitting procedure also provides a principled way for merging the two representations of each vowel. The routine available in MATLAB is based on a Delaunay triangulation of the input data. This in effect provides an estimate of how well each point on the grid is supported by the input data, leading to a weighting scheme for merging axial and coronal datasets. In other words, grid points associated with large-area triangles are considered less reliable, and weighted lower. This avoids a priori specification of the weighting scheme (e.g assign axial slices high weight for tongue root, low weight for predorsum, and vice-versa for coronal); rather, the procedure can adapt automatically to e.g sharply-domed vs. flat tongue shapes. With this scheme, it is also perfectly possible to merge in sagittal data with coronal and axial data if desired.

After conversion of the final radial coordinates back to Cartesian positions relative to the jaw the data is in appropriate form for statistical modelling. Since the complete multispeaker dataset with the most recent version of the surface generation is not quite ready, reports on the results of the N-way procedures will be kept for a later date, and we will confine ourselves here to an illustrative example based on simple PC analysis for one speaker (see Fig. 4). Factor 1 (capturing low back vs. high front) strongly ressembles the frequently encountered "front-raising" factor. It explained an unusually high proportion of the variance, namely about 89%, probably because of the rather small number and unbalanced arrangement of the vowels in the analysis. It is interesting to observe the changes in tongue width associated with Factor 1 in the coronal and axial views. In terms of tongue grooving there is a kind of reciprocal relationship between the poles of Factor 1 in the coronal vs. axial display. Factor 2 explained only about 8% of the variance but is nonetheless clearly crucial for adequate differentiation of the vowel space.

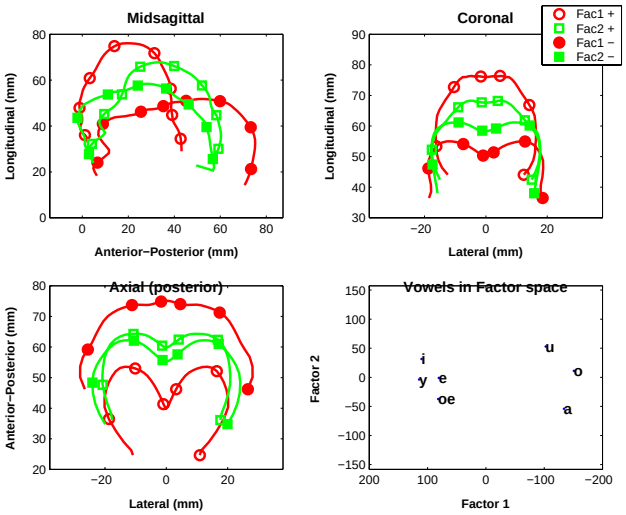

Figure 4:Tongue shapes (+/- 2 s.d) and vowel space associated with first two principal components. Coronal and axial contours run upwards and rearwards respectively from the centre of the tongue.

ACKNOWLEDGEMENTS

Work supported by DFG Ti69/30 and NTT Basic Research Labs. Dr. Axel Wismüller supervised MRI acquisition. Data for 2D/5D comparison acquired in collaboration with Chris Harris and Lucy Ellis.

REFERENCES

[1] P. Badin, G. Bailly, L. Reveret, M. Baciu and C. Segebarth, "Three-dimensional linear articulatory modelling of tongue, lips and face, based on MRI and video images," *J. Phonetics*, vol. 30(3), pp. 533-554, 2002.

[2] J.C. Gower, "Generalized Procrustes Analysis," *Psychometrika,* vol. 40 (1), pp. 33-51, 1975.

[3] R. Harshman, P. Ladefoged and L. Goldstein, "Factor Analysis of Tongue Shapes," *J. Acoust. Soc. Am.* vol. 62, pp. 693-707, 1977.

[4] P. Hoole, A. Wismüller, G. Leinsinger, C. Kroos, A. Geumann and M. Inoue, "Analysis of tongue configuration in multi-speaker, multi-volume mri data," *Proc. 5th Speech Production Seminar*, pp. 157-160, 2000.

[5] P. Hoole, C. Geng and R. Winkler, "Towards a speaker-independent representation of tongue-posturing for speech", in *Speech motor control in normal and disordered speech*, : B. Maassen et al., Eds., pp.138-141. Nijmegen, 2001.

[6] L.R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, pp. 279-311, 1966.

[7] Y. Zheng, M. Hasegawa-Johnson and S. Pizza, "Analysis of the three-dimensional tongue shape using a three-index factor analysis model," *J. Acoust. Soc. Am.*, vol. 113(1), pp. 478-486, 2003.

[8] A. Zierdt, P. Hoole and H.G. Tillmann, "Development of a System for Three-Dimensional Fleshpoint Measurement of Speech Movements," *Proc. XIVth ICPhS*, 1999.