Towards a speaker-independent representation of tongue posturing for speech

Philip HOOLE[1], Christian GENG[2] and Ralph WINKLER[2]
[1]*Institut für Phonetik und Sprachliche Kommunikation, Munich University*
hoole@phonetik.uni-muenchen.de
[2]*Zentrum für Allgemeine Sprachwissenschaft, Berlin*
geng@zas.gwz-berlin.de, winkler@zas.gwz-berlin.de

Abstract. In previous work the three-mode factor analysis technique PARAFAC had given a revealing picture of tongue configurations for vowels derived from both EMMA and mid-sagittal NMRI data. While very parsimonious and elegant, it appeared, however, that the PARAFAC model might be too restrictive to capture speaker-specific effects in consonantal (from EMMA) and non-midline (from NMRI) aspects of articulation. In the current work we are exploring more general statistical models. To make NMRI data amenable to the statistical techniques we outline a jaw-based registration method followed by extraction of intrinsic tongue-surface coordinates on a spherical grid.

1. Introduction

A central question in speech production research is the number and nature of the underlying building blocks that speakers use to organize their articulatory activity; a related question is the extent to which such organizational principles are shared over speakers. The PARAFAC technique of factor analysis [Harshman et al., 1977] has proved useful for addressing these issues. By making strong assumptions about possible speaker-specific features it has been claimed that it can recover organizational principles not directly observable in the raw data. Conversely, if the algorithm fails, this helps to identify more precisely how speakers' articulatory behaviour can differ. PARAFAC is an example of a 3-mode technique of factor analysis; in our case the dimensions correspond to speech items (e.g vowels), articulators, and speakers. This contrasts with standard two-mode analysis where the data are arranged in a two-dimensional array of observations for a set of variables. If a stable solution can be extracted using this three-mode approach, then the problem of rotational indeterminancy inherent in two-mode techniques is avoided. However, the strong constraint on speaker-specific behaviour resides in the fact that all speakers must be assumed to use the same underlying factors. Differences between speakers must be captured in a single multiplicative weight per speaker and factor.

We have already used PARAFAC to investigate two basic kinds of speech data. In Hoole (1999), EMMA data of vowel articulation was analyzed. While basic vowel postures were readily extracted, consonant influences on the vowels proved intractable within the PARAFAC framework; it was necessary to apply a speaker-specific analysis of the residue remaining after extraction of an initial PARAFAC model, in order to arrive at a model of the complete dataset In Hoole et al. (2000) MRI data of vowel articulation was analyzed. Again, the model was successful when applied to the mid-sagittal portion of the MRI data, but there was clear evidence that speaker-specific details of 3D tongue shapes were likely to prove intractable for the algorithm. In view of such problems, and other possible problems with degeneracy of PARAFAC models discussed in the articulatory literature (see e.g Nix et al.'s (1996) reanalysis of a 3-factor solution proposed by Jackson (1988) for Icelandic), it is useful to go into the general mathematical formulation of these statistical models by way of background to the approaches we are now exploring.

PARAFAC can be seen as only one in a broader class of three- or even n-way methods, of which the general notion was given by Tucker (1966). The so-called Tucker3 model of factor analysis reduces the dimensionality of all three modes in the three-way data through a limited set of factors. The three modes are treated symmetrically. Tucker's model is given as

$$\hat{x}_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} \qquad (1)$$

where P, Q and R are the reduced dimensionalities of the three modes. **G** is the so-called core matrix with dimensions P, Q and R, while i, j and k correspond to the indices into the three-dimensional data array. PARAFAC can now be formulated as a special case of Tucker with $P = Q = R$, i.e in the case of three-way data, the Tucker core array of the PARAFAC model can be seen as a cube with elements only on the main diagonal (Kiers, 1991). The PARAFAC model effectively provides only one set of factors, instead of three for the Tucker models. These components belong to all three modes simultaneously, which facilitates the interpretation of results tremendously: Firstly, the core 3-way array **G** has in effect collapsed into a vector; secondly, the solution loses its rotational

indeterminacy.

While the general Tucker model, unlike PARAFAC, is not unique there do still exist various approaches for fitting a reasonable model. For example, Harshman, Lundy and Kruskal (discussed in Kroonenberg, 1992) present the procedure PFCORE for relaxing constraints on the solution. After deriving a PARAFAC solution with orthogonality constraint they suggest using the results from that solution to investigate the nature of the core matrix for a Tucker model. Another strategy is to fit a Tucker model and then rotate the solution according to a specified criterion (Kiers, 1992). Alternatively, there exist procedures setting specific core elements to zero by minimizing the explained sum of squares of the elements set to zero (Kiers, 1998)

Clearly, the down side of all this is that a Tucker model, in addition to being less parsimonious, is also less easy to interpret. This is because the interpretation not only involves the components themselves for all three modes, but also all interactions between these components (as given by the core, see above). Additional complexity could arise from the potential necessity to increase the dimensionality of the design: For example, the dataset in Hoole (1999) was arranged for Parafac analysis as an array of vowels*articulators*speakers. In fact, the vowel mode consisted of 45 combinations of 15 vowels in 3 consonant contexts, and the speaker mode consisted of the 14 combinations of 7 speakers recorded in 2 speech-rate conditions, so it would be conceivable to cast this dataset as a 5-way array, where consonant context and speech-rate also form separate modes.

We are now starting to explore the application of these new approaches to the data used in Hoole (1999) and Hoole et al. (2000), and ultimately aim to determine to what extent speaker-independent and speaker-specific aspects of articulation can be explicitly separated within the framework of a single model, while incorporating non-vocalic and non-midline articulatory behaviour.

2. Material

EMMA data (Hoole, 1999): X/Y coordinates of four fleshpoints on the tongue for 15 vowels of German (all stressable, monophthongal vowels) , spoken in 3 symmetrical consonant contexts (/p, t, k/) at 2 different speech rates by 7 speakers.

NMRI data (Hoole et al., 2000): Vocal tract scans in sagittal, coronal and axial orientations for seven long vowels of German. All scans had a slice thickness of 4mm, and an interslice gap of 1mm. All three volume orientations encompassed the vocal tract completely. Image resolution was approx. 1mm per pixel.

3. Brief overview of previous results

Figs. 1 and 2 show the tongue configurations associated with the factors of a two-factor PARAFAC solution for the EMMA and MRI data respectively. There is broad agreement: Factor 1 captures the contrast between low back and high front; it resembles the factor originally designated “front raising” by Harshman et al., which seems to consistently emerge from studies of this kind. Factor 2 captures variation between low(mid) front to high back. Nevertheless there are differences of detail. For example, for the high, back configuration found with Factor 2, the tongue appears to bunch as it moves back and up in the MRI data, while in the EMMA data the picture is more one of the whole tongue simply moving back and slightly up. Moreover the MRI data makes it clear that this high, back configuration is associated with advancement of the tongue root, whereas by the nature of things this could not be captured in the EMMA data.

The work now in progress aims to apply the range of approaches discussed in the introduction with the following specific points in mind:

Related to the EMMA work: How might speaker-specific features of consonantally-related articulation best be captured? These appeared to be the reason why we could not extend the Parafac model beyond the two factors shown above, although these two factors on their own clearly did not capture all coarticulatory influences of flanking consonants on the vowels. A preliminary reanalysis of these data using the Tucker model resulted in an unevenly shaped core matrix **G**: factor versus fit plots suggested the use of 3 factors in the vowel mode and in the articulator mode, but only one or two in the speaker mode. However, we are still in the process of investigating to what extent off-main-diagonal core elements may provide a useful indication of the nature of speaker-specific effects.

Related to the MRI work: Inspection of selected coronal and axial contours had made it appear unlikely that the original Parafac framework would be able to handle the differences between speakers when extended to 3-dimensional tongue data. However, there remained the task of actually extracting the 3D coordinates of operationally homologous points over the whole tongue surface for all speakers and merging the data from all three volume orientations.

4. Steps in MRI analysis

4.1. Determination of tongue contours in each MRI slice

This was done by interactively combining contours derived from automatic detection using a fixed intensity threshold criterion at the tissue-air interface (using an edge-detection algorithm to ensure consistent choice of

threshold level over volumes and spekaers), with manually placed points in other regions. In cases where it was difficult to 'separate' the tongue from the hard palate (e.g for high front vowels), complete contours of the hard palate taken from recordings of low vowels were superimposed (an analogous procedure was followed for contact between tongue and pharyngeal wall in low, back vowels). For every slice a sufficient number of points was defined to allow accurate reconstruction of a complete contour using spline interpolation.

4.2. Jaw-based registration

The next main preparatory step was to extract sufficient information on the jaw to allow rotation and translation of all tongue data to a common jaw position.

This served two purposes: First of all, it helped to compensate for (usually only slight) differences in position of the subject when recording the same vowel at different volume orientations (usually separated by an interval of several minutes). Secondly, the aim was to base the statistical models on intrinsic tongue configuration first, with the possibility for adding in information on jaw position later if desired. Accordingly, this necessitated mapping all data to a constant jaw orientation (the close jaw position for /i/ was chosen as the reference).

First the coronal and axial volumes were resliced so that they could also be viewed as sagittal volumes. Then, in every sagittal slice the most inferior point of the the low intensity region corresponding to the jaw bone was tracked through as many slices as possible (this point is indicated by the white slanting arrow in Fig. 4a for a midsagittal slice). The points extracted in this way were used as input to a generalized Procrustes algorithm (Gower, 1975) to obtain rotation matrices that would best map these sets of landmarks to a common reference. An example of a mandibular contour derived in this way is included in Fig. 3.

4.3. Extraction of points on the tongue-surface

For the actual statistical analyses, a fixed number of operationally homologous points from the tongue surface is required. These were defined as follows: First of all, a point representing the centre of the tongue was defined. To do this a line was drawn horizontally in a posterior direction from the point at the top edge of the high-intensity region of bone marrow in the jaw bone (this is illustrated by the black arrow in the midsagittal section of Fig. 4a). This region was always well-defined. The centroid of the tongue was then calculated in the midsagittal slice using the tongue coordinates above this line. This point was used as the origin of a system of spherical coordinates; in other words the values actually used as input to the statistical procedures were the distance from this origin to the point of intersection of a line at one of the chosen combinations of azimuth and elevation with the tongue surface. Parts of this grid are illustrated in the sagittal, coronal and axial slices shown in Fig. 4a-c (the coronal and axial slices shown are those in which the origin of the spherical coordinate system lies).

To find the points of intersection, the tongue contours of each slice were converted to a fixed number (100) of equidistant points per contour; at each of these points, spline interpolation was used across slices to in turn reduce the spacing between contours from 5mm to 1mm.

By taking points from adjacent contours, polygons were defined to allow a representation of the tongue surface to be generated. This is illustrated in Fig. 3 for a sagittally oriented volume.

The points on the dense set of interpolated contours were used to estimate the distance from origin to tongue surface at each of the spherical coordinates actually required.

Unlike the approach followed in Badin et al. (2000) and Engwall (2000), we are not aiming for a complete tongue reconstruction; rather, we simply want to sample a sufficient number of points on the tongue surface to capture all systematic vowel-induced changes in configuration. Merging of the data from the three volume orientations will be done at this stage, i.e the volume providing the most reliable tongue-surface estimate for a given spherical coordinate will be used.

The complete set of values input to the statistical procedures will also include the coordinates of the origin relative to the position on the jaw marked by the start of the horizontal black arrow in Fig. 4a, as well as the length of this arrow itself, i.e the horizontal distance from the posterior midline edge of the jaw to the tongue root. It was felt that this measure could reflect rather directly the influence of the GGP, which undoubtedly is one of the major influences responsible for the tongue configuration as a whole.

5. The state of play

At the time of writing, the donkey-work of extracting the required data for all vowels, subjects and volume orientations is still in progress. In parallel with this, we are testing the implementation of the statistical algorithms outlined in the introduction. Shortly these two strands will meet.

References

Badin, P., Borel, P., Bailly, G., Reveret, L., Baciu, M. & Segebarth, C. (2000). Towards an audiovisual virtual talking head: 3D articulatory modelling of tongue, lips and face based on MRI and video images. *Proc. 5th*

Speech Production Seminar, 261-264
Engwall, O. (2000). Replicating three-dimensional tongue shapes synthetically. *TMH-QPSR 2-3/2000*, 1-12.
Gower, J.C. (1975). Generalized Procrustes Analysis. *Psychometrika* 40 (1), 33-51.
Harshman, R., Ladefoged, P., and Goldstein, L. (1977). Factor Analysis of Tongue Shapes. *J. Acoust. Soc. Am.* 62, 693- 707.
Hoole, P. (1999). On the lingual organization of the German vowel system. *J. Acoust. Soc. Am.* 106(2), 1020-1032.
Hoole, P., Wismüller, A., Leinsinger, G., Kroos, C., Geumann, A. & Inoue, M. (2000). Analysis of Tongue Configuration in Multi-speaker, Multi-volume MRI Data. *Proc. 5th Speech Production Seminar*, 157-160
Jackson, M. T. T. (1988). Analysis of tongue positions: Language-specific and cross-linguistic models. *J. Acoust. Soc. Am.* 84, 124-143.
Kiers, H.A.L. (1991). Hierarchical relations among three-way methods. *Psychometrika*, 56(3), 449-470.
Kiers, H.A.L. (1992). Tuckals core rotations and constrained tuckals modelling. Statistica Applicata, 4(4), 659-667.
Kiers, H.A.L. (1998). Three-way simplimax for oblique rotation of three-mode factor analysis core to simple structure. *Computational statistics and data analysis*, 28, 307-324.
Kiers, H. A. L., ten Berge, J. M. F & Bro, R. (1999). PARAFAC2-Part I. A direct algorithm for the PARAFAC2 model. *Journal of Chemometrics* 13, 275-294.
Kroonenberg, P.M. (1992). Three-mode component models. A survey of the literature. Statistica Applicata, 4(4), 619-633.
Nix, D. A., Papçun, G., Hogden, J. & Zlokarnik, I. (1996). Two cross-linguistic factors underlying tongue shapes for vowels. *J. Acoust. Soc. Am. 99*, 3707-3718.
Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279-311.

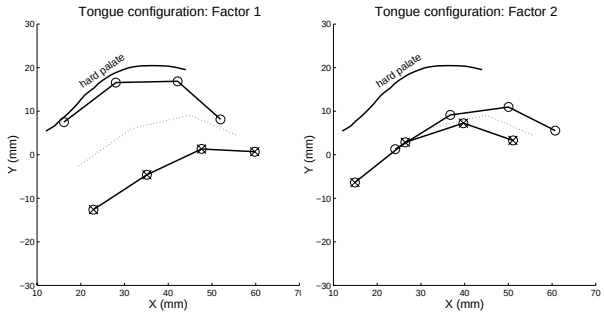


Fig. 1. Tongue shapes for first factor (left) and second factor (right) of Parafac analysis of EMMA data

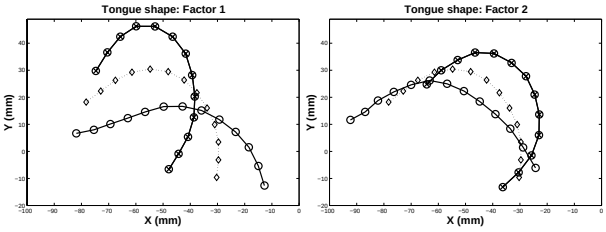
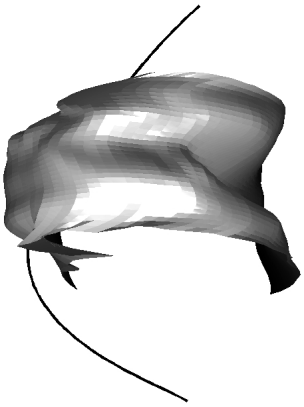


Fig. 2. Tongue shapes for first factor (left) and second factor (right) of Parafac analysis of NMRI data



←Fig. 3. Tongue surface for /i/ from sagittal NMRI volume. Jaw contour also shown.

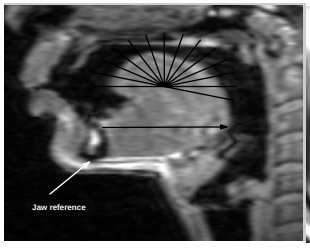


Fig. 4a. Midsagittal scan of /u/. For explanation of arrows and grid see text.

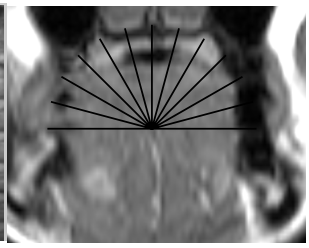


Fig. 4b. Coronal scan of /u/, with measurement grid.

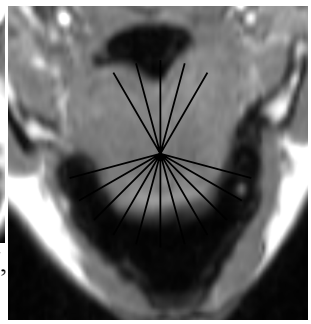


Fig. 4c. Axial scan of /u/ with measurement grid