ANALYSIS OF TONGUE CONFIGURATION IN MULTI-SPEAKER, MULTI-VOLUME MRI DATA

Phil Hoole[1], Axel Wismüller[2], Gerda Leinsinger[2], Christian Kroos[1,3], Anja Geumann[1], Michiko Inoue[1]

[1] *Phonetics Institute, Ludwig-Maximilians-Universtität, Munich*
[2] *Radiology Department, Klinik Innenstadt, Ludwig- Maximilians-Universtität, Munich*
[3] *ATR HIP, Kyoto.*

email: hoole@phonetik.uni-muenchen.de

ABSTRACT

MRI data of German vowels and consonants was acquired for 9 speakers. In this paper tongue contours for the vowels were analyzed using the three-mode factor analysis technique PARAFAC. After some difficulties, probably related to what constitutes an adequate speaker sample for this three-mode technique to work, a stable two-factor solution was extracted that explained about 90% of the variance. Factor 1 roughly captured the dimension low back to high front; Factor 2 that from mid front to high back. These factors are compared with earlier models based on PARAFAC. These analyses were based on midsagittal contours; the paper concludes by illustrating from coronal and axial sections how non-midline information could be incorporated into this approach.

1. INTRODUCTION

A large body of MRI data has been acquired with the aim of throwing further light on issues raised in experiments already completed using data from electromagnetic midsagittal articulography (EMMA). Briefly, the MRI data consist of complete coronal, axial and sagittal volumes of the long German vowels /i, e, y, ø, a, o, u/ as well as of the alveolar consonants /t, s, n, l, ʃ/.

The first general issue being addressed involves extraction of vocal tract area functions for both vowels and consonants to assist in interpretation of patterns of articulatory variability observed in EMMA experiments in terms of area-function variation.

The second general issue (which will be the focus of the present paper) involves extraction and analysis of tongue rather than complete vocal tract configurations. This is being carried out on the background of the EMMA experiment presented in Hoole (1999) in which three-mode factor analysis (PARAFAC; cf. Harshman et al., 1977) was used to identify underlying patterns of tongue shapes in a multi-speaker data-set of German vowels. This study confirmed PARAFAC's potential for giving a very parsimonious representation of multi-speaker data; moreover the two basic vocalic factors extracted appeared phonetically and physiologically plausible.

Nevertheless, for these EMMA analyses the concern must remain that the nature of the extracted factors may be distorted by the paucity of direct information on pharyngeal and tongue-root configuration. In addition, there is the possibility that a different partitioning of the data observable at the articulatory surface into underlying behavioural "building-blocks" may occur if the inherently 3D nature of the tongue is taken into account. Following analysis of midsagittal contours, the analyses are now just reaching the stage where the 3D information can be incorporated directly. There was in fact a very good methodological reason for not launching straight into a full-blown PARAFAC analysis of 3-dimensional data: Our previous experience with EMMA data had shown that deriving a stable PARAFAC model could prove tricky even in cases where a priori no problems were expected (for example, in Hoole, 1999, it proved easy to derive a model for vowels spoken in pVp (and kVk) context, but not in tVt context).

Thus it seemed important as a first step to demonstrate that the MRI data were indeed amenable to analysis along traditional (i.e in this case midsagittal) lines. As in the past it again turned out that an apparently simple task would be surprisingly recalcitrant, with the difficulties on the way providing some useful insight into what can be expected from this modelling procedure.

The final aim of the work outlined here is to provide a background for most effective choice of sensor locations for the incipient 3D EMA system (i.e EMMA without the midsagittal constraint) reported on elsewhere (Zierdt et al.,1999; Zierdt et al., this meeting), and to give an idea of plausible changes in sensor alignment that can be expected to occur as the tongue changes shape, since the 3D EMA is able to measure 2 sensor orientation angles in addition to the 3D spatial coordinates.

2. RECORDING PROCEDURES

To date, 9 speakers have been recorded (8 male, 1 female); all are phonetically trained. Seven speakers recorded the 7 vowels and 5 consonants given above, two speakers recorded only the vowels.

Complete sets of coronal, axial and sagittal scans were performed. Scans for each volume orientation were performed for all target sounds before proceeding to the next volume orientation. Thus there was an interval of several minutes between utterances of the same sound in different orientations. Subjects were prompted to start production of the target sound and initiation of the acquisition was done a couple of seconds later allowing a short stretch of speech to be recorded over the talkback system uncontaminated by the noise of the scanner. If possible, subjects also prolonged phonation a few seconds beyond termination of the acquisition.

All volumes were recorded with a slice thickness of 4mm. For the first 6 subjects all volumes had a 1mm interslice gap and consisted of 23 slices. Such sequences took about 20s to acquire. For the three most recent subjects, sagittal volumes had only 13 slices and no interslice gap, giving a corresponding reduction in acquisition time.

With these settings all three volume orientations encompassed the complete vocal tract.

Acquisition details: Siemens Magnetom Vision, 1.5 Tesla, T1 weighted FLASH sequence, TE=4.1ms, TR=182.9ms. (In addition, special-purpose sequences were used for imaging the condyle (for use in jaw movement analysis to be reported on elsewhere), and we are also collecting scans of the subjects' dental impressions for insertion into the edentulous MRI volumes.)

Notes on the speech material: Only the long monophthongal vowels of German were selected for recording in view of doubts as to whether subjects could produce artificially prolonged short vowels appropriately. The long vowel /ɛ:/ (as in "bäten") was omitted in view of its now rather marginal phonemic status for many speakers even though it would actually be a very useful vowel in terms of mapping out the complete vowel space.

3. ANALYSIS

As mentioned in the introduction it was considered to be an essential first step to show that a PARAFAC-based model could be extracted from midsagittal data.

The general procedure followed quite closely that used in Nix et al.'s (1996) reanalysis of Harshman et al.'s original radiographic data. Rather than measuring the tongue based on its intersection with a predefined anatomical grid (as done in Harshman et al.) each tongue contour was captured by 13 equally-spaced points from the tip to the base. The 13 pairs of xy-coordinates for these points were treated as the coordinates of a set of 13 "virtual pellets". Nix et al.'s argument in favour of this procedure over the predefined grid approach was that the latter does not allow horizontal movements to be directly captured. We think this is a fair point; moreover, the "virtual pellet" approach was attractive because our previous work used true fleshpoint data (from EMMA). Of course, this approach also has clear hazards: The tongue may not extend and contract uniformly; also, the procedure requires one to designate one point on each tongue contour as the tip of the tongue (constituing the frontmost virtual pellet). But it would be extremely surprising if a truly identical fleshpoint is chosen for each vowel analyzed. Discrepancies of one or two mm are probably inevitable. At the other end of the tongue contour this kind of uncertainty is possibly even more acute. After some experimentation we chose to truncate the tongue contour at the vertical location adjacent to the tip of the epiglottis.

Given that the first set of analyses was to be carried out on midline data then ideally one need do no more than consider the sagittally oriented volumes. For obvious reasons, however, contours obtained on this basis could give a considerably distorted picture of the midline contour of the tongue. There may well be no sagittal slice precisely aligned on the midline; even if this could be resolved by interpolation and realignment, the thickness of the slices could still result in unclear contours where sharp tongue grooving occurs (for examples of this refer forward to the intersubject comparison of axial and coronal sections given at the end of the paper in Figs. 7 and 8).

Despite these clear disadvantages, the first stage carried out was to extract from the sagittal volume data a contour as close to mid-sagittal as possible. The first reason was that the sagittal view was the easiest one in which to determine the locations of the tip of the tongue and the tip of the epiglottis. Secondly the midsagittal slice was used to extract a contour of hard palate, rear pharyngeal wall and spinal region. These contours in turn formed the basis for adjusting the data from the the different vowels (of a given speaker) to correct for slight changes in head position. Fig. 1 gives an example of this kind of raw trace.

The plausibility of the extracted tongue contour was then crosschecked by overlapping data from the other volume series: For the oral region from the corresponding coronal volume, for the pharyngeal region either from the coronal volume mapped to an axial grid, or from the axial volume itself.

After overlaying and aligning the data derived from the sagittal, coronal (and axial) volumes, the tongue contour derived from the sagittal volume was manually adjusted if data from the other volume(s) consistently indicated a more plausible path for the contour. For example, this was quite frequently the case for high front vowels due to sharp tongue grooving in the pharyngeal region.

Following this alignment and correction procedure, the raw contour data was spline-interpolated up to a much higher spatial resolution as an intermediate step to form a basis for then reducing the contour to 13 equally spaced points (cf. Nix et al., 1996).

By way of example, Fig. 2 shows the resulting tongue contours for 3 vowels of one speaker.

The values of these 26 (=13*2) articulatory parameters for the seven vowels and 9 speakers formed the input to the PARAFAC algorithm - after a final preprocessing step in which the means over all vowels for each articulatory parameter and speaker were first subtracted, so that the PARAFAC algorithm in fact deals with *deviations* from mean articulatory positions.

4. RESULTS

Attempts to extract a stable PARAFAC solution taught us (again) a lesson we had learnt in the analysis of EMMA data given in Hoole (1999b), namely that it can be surprisingly difficult to obtain a reliable solution even in cases where no particular difficulty was expected.

4.1 The problem of an adequate speaker sample

We first ran the PARAFAC algorithm after 8 speakers had been analyzed and were disappointed to find that no solution could be obtained without introducing orthogonality constraints (see Hoole, 1999, for discussion and further references to the criteria used to assess the reliability of PARAFAC solutions).

The failure to obtain a solution could mean that the attempt is being made to extract too many factors. However, our failed attempts involved 2-factor models, and it seemed highly unlikely that this kind of vowel data could involve only one factor. The other reason for the failure could be that speakers do not conform to the very restrictive PARAFAC model for capturing speaker-specific differences - involving a single multiplicative weight for each speaker and factor.

At this juncture one approach might have been to explore the possibilities offered by the PARAFAC2 model for relaxing constraints on the nature of the subject-specific behaviour (cf. Geng et al., this meeting, for further discussion of this model, and references to the original formulation of the model by Harshman). This might well be fruitful in a dataset such as that of Hoole (1999) where a consonantally-related factor was identified, whose mapping to subjects' fleshpoints was however outside the scope of the basic PARAFAC model.

In the present case it turned out that a more conservative approach was still feasible. Data for the ninth speaker became available much later than for the other speakers; when this data was included there was suddenly no problem at all in extracting a stable two-factor solution. Below we will examine the solution in terms of the tongue configurations themselves. Before doing so it is worth noting that within the PARAFAC framework there is probably a perfectly rational but nonetheless rather instructive reason why the state of affairs should abruptly change. Consider the speaker weights for the two factors (Fig. 3): The ninth speaker was MH. The point to notice is that he is located at one of the extreme positions in the Factor1/Factor2 space. In other words, introducing this speaker into the analysis introduces a relatively novel combination of Factor1/Factor2 speaker weights. In their original paper Harshman et al. emphasize that a range of combinations of speaker weights is a necessary condition for the PARAFAC algorithm to fulfil its potential to resolve the rotational indeterminancy inherent in standard two-mode factor analysis models.

The sensitivity of the analysis to having a sample of speakers whose behaviour exhibits sufficient variety, but with the nature of the variety still being consistent with the model is undoubtedly a problem, since although PARAFAC has in several studies given acceptable models with only half a dozen speakers, it is by no menas clear what sample size would be required to make extraction of a stable solution more than just a hit-and-miss affair.

After this preamble on the difficulties encountered, we turn to consideration of the solution itself.

4.1 The two-factor model

The two-factor solution, which was the only one meriting detailed consideration, explained about 87% of the variance (with an RMS error of about 2.2 mm). This is somewhat higher error than most previous studies with two-factor models have encountered.

The tongue shapes related to the two factors are shown in figs. 4 and 5. One of the main points of interest for us was in determining to what extent the solution would resemble that found using EMMA in Hoole (1999) for German vowels, with about half the subjects in common. At first sight the similarities are quite substantial: As in the earlier paper Factor 1 captured variation from high front to low back tongue constriction, and Factor 2 from low (or mid) front to high back.

There were some noticeable differences however. In Hoole (1999) for Factor 1 all fleshpoints moved forward as they moved up. In the present study this does not hold true for the frontmost virtual pellets. In the MRI data a picture of anterior-posterior compression of the tongue comes out more strongly. For Factor 2, although the overall displacement of the tongue appears very similar in both studies, the MRI study reveals a stronger accompanying change in tongue shape (i.e the tongue becomes more bunched as it moves back (and up). Moreoever, the overall high back tongue shape is actually linked with more advanced tonge root (also found in the factor Harshman et al. refer to as “back raising”, though our Factor 2 differs from their factor in that the overall retraction component is much more in evidence) - a fact that could not of course be extracted from the EMMA data.

These differences of detail between the two studies may again be partly attributable to the uncertainties in the virtual pellet method, but the availability of the pharyngeal/tongue root information in the MRI data has probably the bigger role to play.

As a consequence of the above differences, the orientation of the vowels in the vowel space in the present study is somewhat different from that found in the EMMA study (and also not identical to either the original Harshman study or the Nix et al. reanalysis; cf. Fig. 6). Roughly speaking, one could say that the new vowel space is rotated about 30 degrees counterclockwise relative to the old one. Curiously, the present arrangement does have some similarity with one version of the traditional vowel chart, with more open front vowels being more “retracted” than close ones, and /u/ not located as “high” as /i/.

5. OUTLOOK

In this concluding section we will briefly illustrate how the three-dimensional shape of the tongue might be incorporated into the present approach. Ultimately, an approach linking the three-dimensional MRI volume data to a more physically oriented model of the tongue as a 3D deformable structure would be desirable. For the present, however, in view of the difficulties that repeatedly become apparent in following a multi-speaker approach through to an acceptable conclusion, we envisage a much more modest approach in which small amounts of information are incrementally added to the dataset on which the model is to be based. In the present case, the next step will be to apply some fairly low-dimensional curve-fitting to the tongue cross-section at each of the virtual pellet positions and, for example, to incorporate coefficients capturing degree of convexity or concavity at each of these fleshpoints into the modelled data (cf. Stone & Lele, 1992).

Figs. 7 and 8 give an idea for our 9 subjects of the kind of contours that will have to be contended with. Fig. 7 shows tongue surface contours taken from axial sections of the vowel /i/ at a slice location about halfway between the tip of the epiglottis and the uvula, while Fig. 8 shows coronal sections for the vowel /a/ taken below the highest point of the palatal vault.

Considering the axial sections first, it will be observed that all subjects show the grooving typically observed for /i/ at this location, but that depth and sharpness of the groove vary rather substantially. Thus, for the sake of argument, this kind of shape information might be systematically related to strongly negative (/i/-like) values of Factor 1. If, in turn, speaker-specific aspects of the grooving (such as depth) covaried with other aspects of their behaviour with respect to Factor 1, then the present model might already be capable of capturing non-midline tongue shaping. Speaker KH, for example, who was observed in the raw data to show rather restrained midsaggittal tongue advancement in the tongue-root region, shows the shallowest groove and also the lowest speaker weight for Factor 1. However, inspection of further speakers makes it clear that no simple generalization exists between tongue-groove pattern for /i/ and the pattern of speaker weights in the current model. This impression is reinforced by consideration of the coronal patterns for /a/. Perhaps the most typical pattern is a small, sharp groove, but the sharpness of the groove still varies considerably, and one speaker (SR) even shows the opposite pattern of concavity/convexity to the other speakers. On balance it appears quite likely that incorporation of this cross-sectional information will make it necessary to have recourse to a model of the PARAFAC2 variety (already mentioned briefly regarding consonantal articulation) in view of the greater flexibility it allows in the mapping between underlying factors and individual speakers’ fleshpoint behaviour.

ACKNOWLEDGEMENTS

Work supported by German Research Council grant DFG Ti 69/31-1. The MATLAB toolbox of Mark Tiede (ATR-HIP) was a great source of inspiration for MRI volume analysis and display.

REFERENCES

Geng, C. & Mooshammer, C. (2000). “*Modeling the German stress distinction*”. Proc. 5th Speech Production Seminar (this volume).

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.

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.

Stone, M. & Lele, S. (1992). “*Representing the tongue surface with curve fits*”. Proc. ICSLP 92, pp. 875-878.

Zierdt, A.; Hoole, P.; Tillmann, H.G. (1999). "*Development of a System for Three-Dimensional Fleshpoint Measurement of Speech Movements*". Proc. XIVth Int. Cong. Phon. Sci., pp. 73- 76.

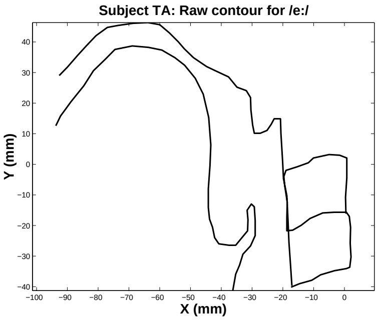

Fig. 1: Example of raw midsagittal tracing of tongue, palate and spinal region

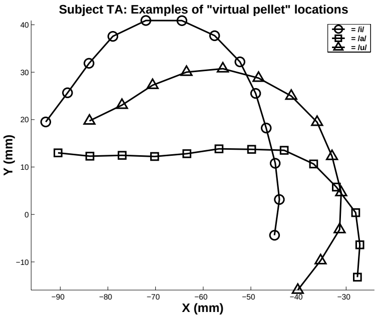

Fig. 2: Examples of the tongue contours for 3 vowels given by 13 equidistant points on each contour./i/=circles, /a/=squares, /u/=triangles. Same data orientation as in Fig. 1

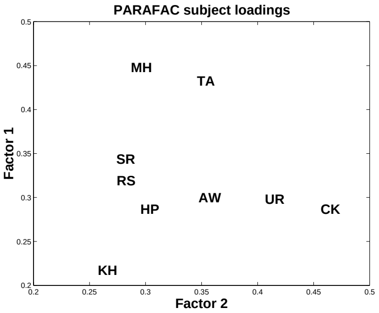

Fig. 3: Subject loadings for the two-factor PARAFAC solution

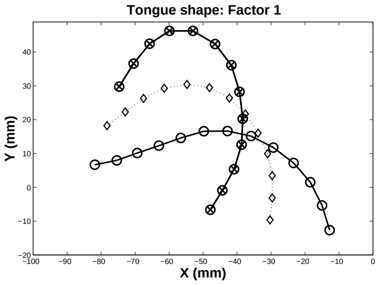

Fig. 4: Factor 1 of the two-factor PARAFAC solution. Deviations from mean tongue configuration (dotted line with diamonds) obtained by setting the factor to +/- 2 s.d. Positive deviation: Unfilled circles. Data orientation as in Fig. 1.

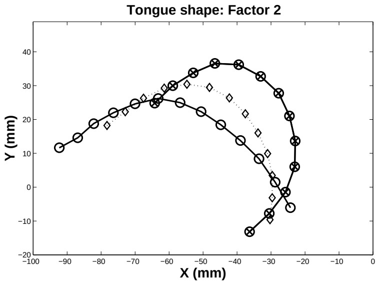

Fig. 5: Tongue shapes associated with Factor 2. Other details as in Fig. 4

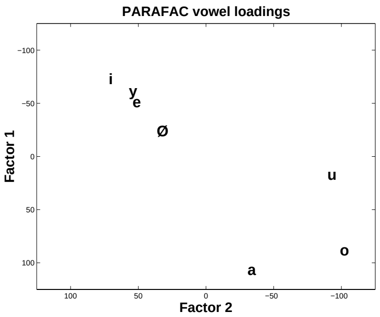

Fig. 6: Distribution of the German vowels in the factor space. Both axes have been reversed to give a more "traditional" orientation.

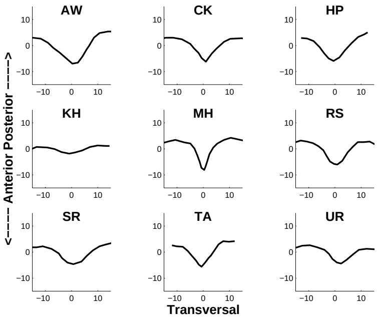

Fig. 7: Example of tongue contours for /i/ extracted from axial sections of all 9 subjects. Section taken midway between tip of epiglottis and uvula.

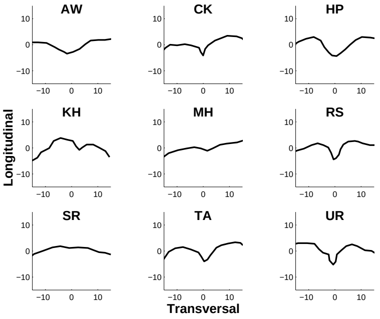

Fig. 8: Example of tongue contours for /a/ extracted from coronal sections of all 9 subjects. Section taken below highest point of palatal vault..