Issues in the acquisition, processing, reduction and parameterization of articulographic data.

Philip Hoole, Munich

(This report is available via Internet from http://www.phonetik.uni-muenchen.de)

This report will give an overview of the chain of procedures involved in acquiring reliable articulographic data, up to the point at which parameters have been extracted and are available for statistical processing. The topics to be covered are divided into two groups. The first group involves procedures that are common to virtually any articulographic investigation. The second group uses the DFG project "Deutsche Vokale" as a specific example to illustrate procedures for extracting parameters from the raw articulographic signal streams.

I. Basic Procedures

This section will cover calibration, preparation of the subject, sensor arrangement, quality control during the experiment, post-processing of articulatory data, synchronization with acoustic data.

1. Calibration

First of all, it may be helpful to recall the principle by which an EMA system measures movement in the two-dimensional plane. The helmet worn by the subject supports three transmitters that are located at the vertices of an equilateral triangle. The first step is to determine from the strength of the induced electromagnetic field the distance from each sensor to each transmitter. The second step involves the application of a geometrical algorithm to convert for each sensor the three distance measurements to two-dimensional x/y coordinates in the measurement plane. Thus the accuracy of the x/y measurements depends in the first instance on having an accurate conversion from induced voltage to transmitter-sensor distance. For articulographic systems the voltage-distance (V-D) relation is known to be roughly of the form V=k*1/D**3 (k being an amplification factor). The exponent in this equation is actually slightly less than 3 and needs to be determined empirically. With the original calibration procedures provided by the manufacturer this was not possible. We were able to show some time ago (Hoole, 1993) using fairly rudimentary methods that for this particular equipment a clear improvement in accuracy could be obtained by determining the exponents individually for each sensor-transmitter combination. The manufacturer thereupon developed a device allowing each sensor to be placed systematically at a wide range of known positions in the measurement field, giving for each sensor-transmitter combination a wide range of voltage-distance value pairs, this in turn allowing the exponent in the above equation to be determined by regression. For better understanding of the following remarks the illustrations of the calibration device should be referred to (Figs.1 and 2).

Briefly, a magazine holding five sensors is locked into a rotatable arm at a known distance from the centre of the measurment field, this being the centre of rotation. The basic radii available are 8 cm, 4 cm and (for special purposes, see below) 0 cm. The arm can be rotated through 360 deg., with measurements being made at 24 pre-defined angular positions (i.e at angular incrememts of 15 deg.). Since the magazine can be locked into the rotating arm at 5 different positions 1 cm apart, 5 modifications of each of the 3 basic radii are possible, giving a potential maximum of up to 3*5*24 = 360 measurement positions per sensor. Since the calibration device must be located by hand at each of these positions it is usually much too time-consuming to use them all. At the final stage of the

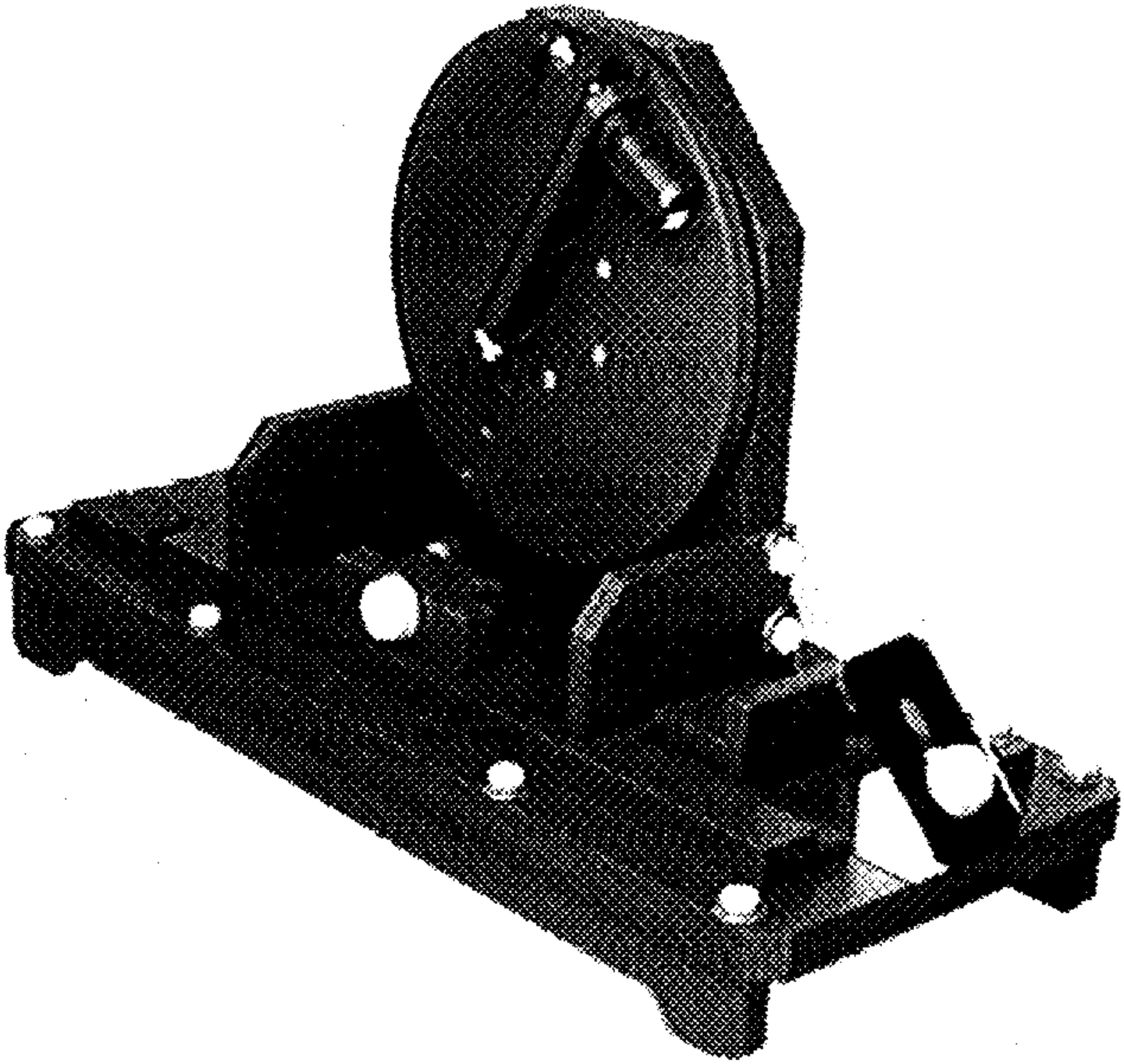

*Fig. **1** Original photo of the Mkal calibration device (Carstens Medizinelektronik, Göttingen).*

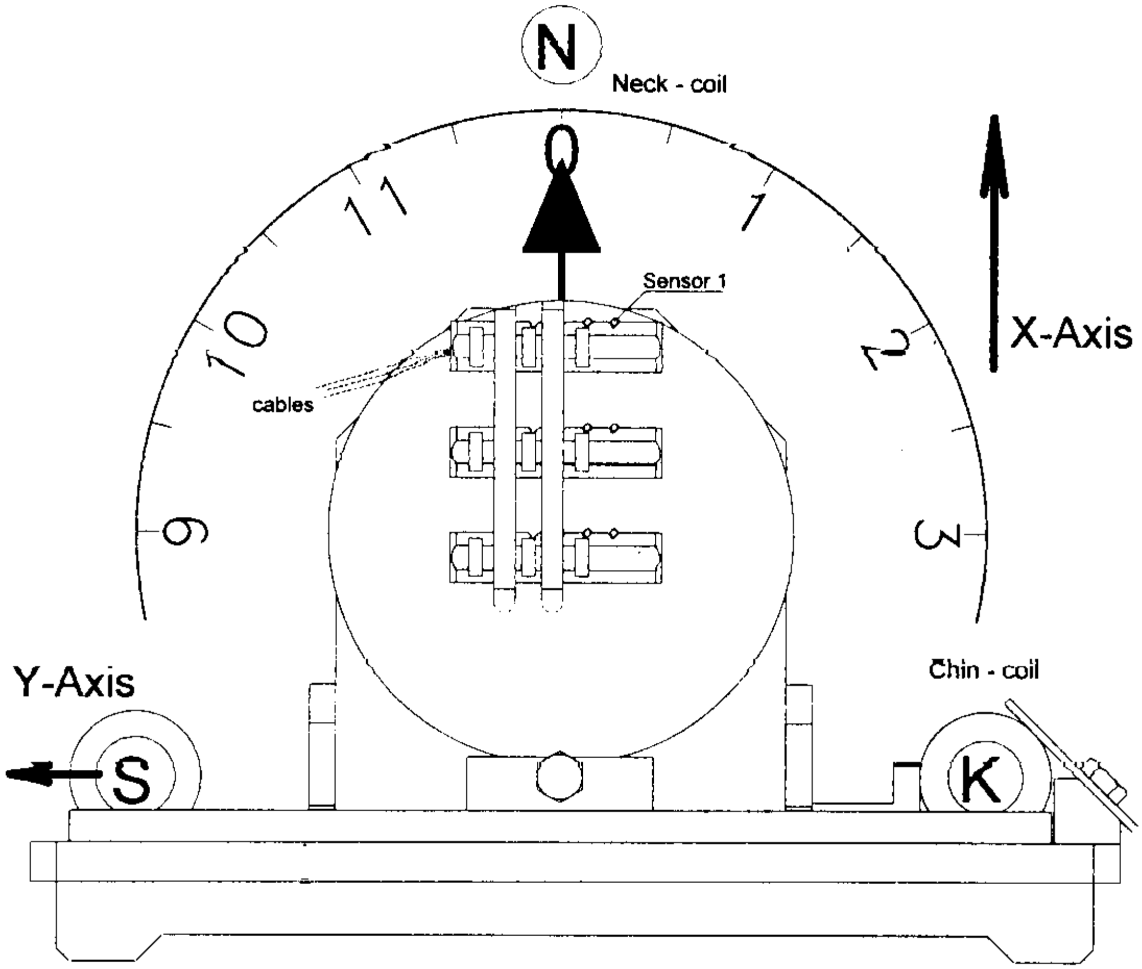

Fig. **2** *Schematic view of calibration device. The helmet is located on the device with the rotating disk in the mid-sagittal plane. The transmitter locations correspond to the circles marked K, S and N. The 3 possible radial positions of the sensor-magazine are shown. From bottom to top they are 0, 4 and 8 cm respectively.*

calibration procedure we use a selection of 144 positions chosen to give the most useful distribution of data in the measurement field.

The Mkal device has proved invaluable for providing a systematic test of system function. However, in order to derive maximum benefit from it we were forced to the conclusion (after considerable experimentation) that a rather painstaking multi-stage calibration procedure was required.

The main problem that confronted us was that the empirical determination of the V-D relation gave excellent results (residual errors of around 0.25 mm) in the centre of the measurement area, which we can define for present purposes as the circular area of 4 cm radius around the centre, but accuracy deteriorated rather noticeably at the edges of the measurement field (again, in practice, meaning in the vicinity of the 8 cm radius about the centre) the error frequently going up to the order of 1-2 mm. We first suspected, based on results reported by Perkell et al. for the MIT system, that this might be due to the fact that the exponent is not in fact constant, but itself depends to a certain extent on the distance from the transmitter (the expectation being for smaller values close to the transmitter and higher values further away). However, a considerable amount of experimentation along these lines did not result in any worthwhile improvement. This led to the suspicion that the optimal voltage-distance conversion depends not just on absolute distance per se, but also on the direction to the sensor as seen from the transmitter, in other words on the location of the sensor in x/y space. This is readily conceivable if, for example, the geometry of the transmitter arrangement deviates from the perfect equilateral triangle, or if, due to idiosyncracies in the electromagnetic field, the V-D relationship differs slightly for a line directly from transmitter to centre of the measurment field, compared, say, to a line 10 or 20 degrees on either side of this. If this is the case then measurement errors will remain no matter how sophisticated one tries to make the mathematical model of the V-D relationship. (Or strictly speaking, a completely different kind of model may be required in which the V-D function is itself a function of x/y position. This in turn would have the drawback that the results could not be captured in look-up tables of realistic size, this being an important aspect contributing to the design feature of real-time performance.)

The solution we eventually chose to this problem is conceptually actually very simple. We recall that the calibration device allows each sensor to be placed in a large number of positions within the measurement field. At each of these points the difference between true and measured position can be determined. If these error vectors show systematic behaviour, i.e change slowly over the measurement field, then it is a fairly straightforward matter to use this information to apply a correction factor to the x/y coordinates of the articulatory movements that have been measured. "Slowly" in the previous sentence means relative to the couple of centimetres of movement typically shown by fleshpoints during articulation. By and large, this has turned out to be the case. Figs. 3 and 4 show maps of the error vectors obtained for 10 sensors in a typical calibration trial. Note that the magnitude of the vectors has been multiplied by 10 to increase visibility. It will be observed that the pattern of the errors is quite similar for all sensors. For example, all sensors show a similar kind of increased error in the lower left corner of each part-figure (error vectors pointing down and to the right). Thus a large part of the remaining error in the system appears to be systematic rather than random.

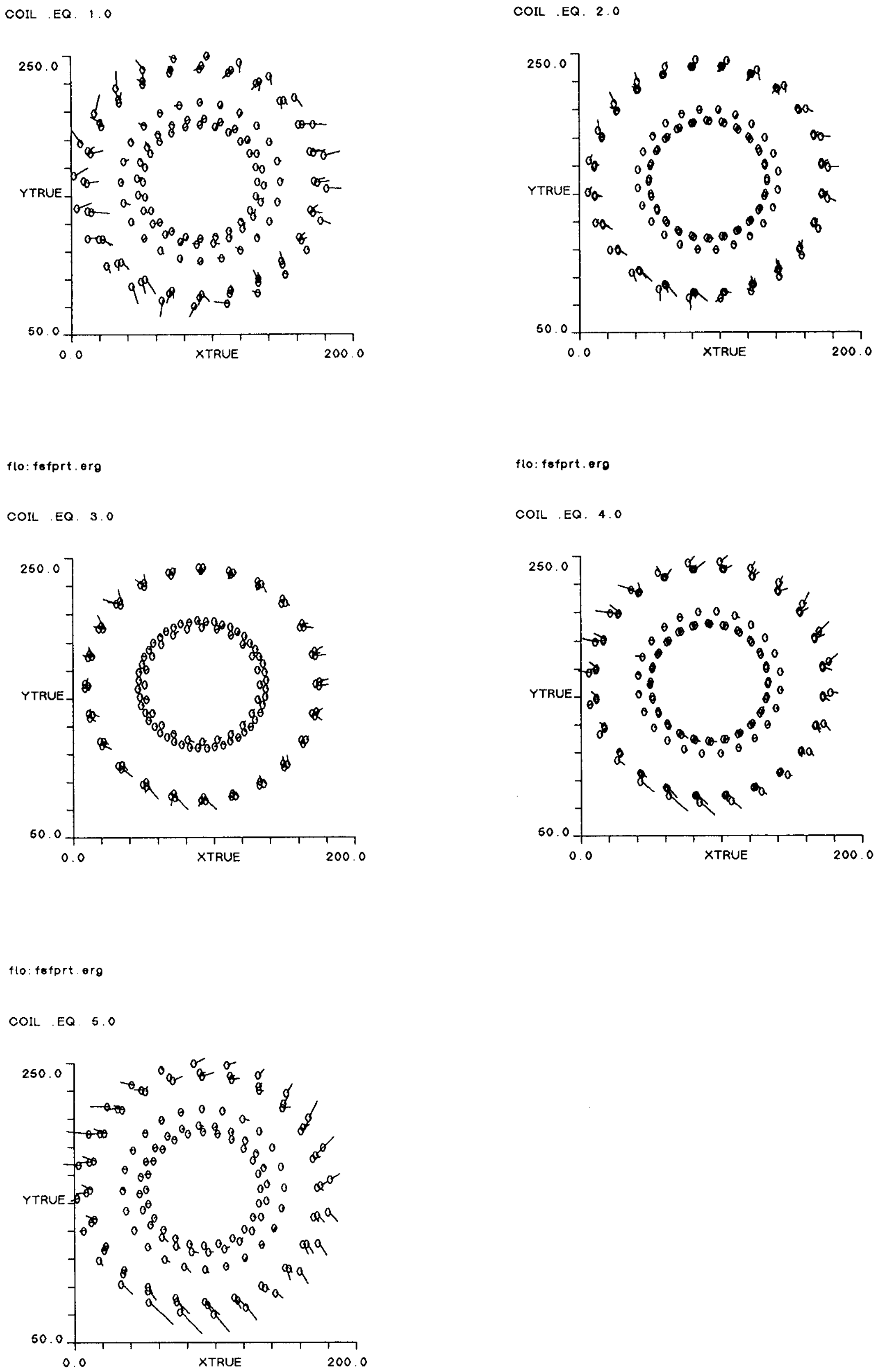

Fig. 3 Error maps of sensors 1 to 5. True position indicated by small circles. Coordinate axes in mm. Error vector length multiplied by 10.

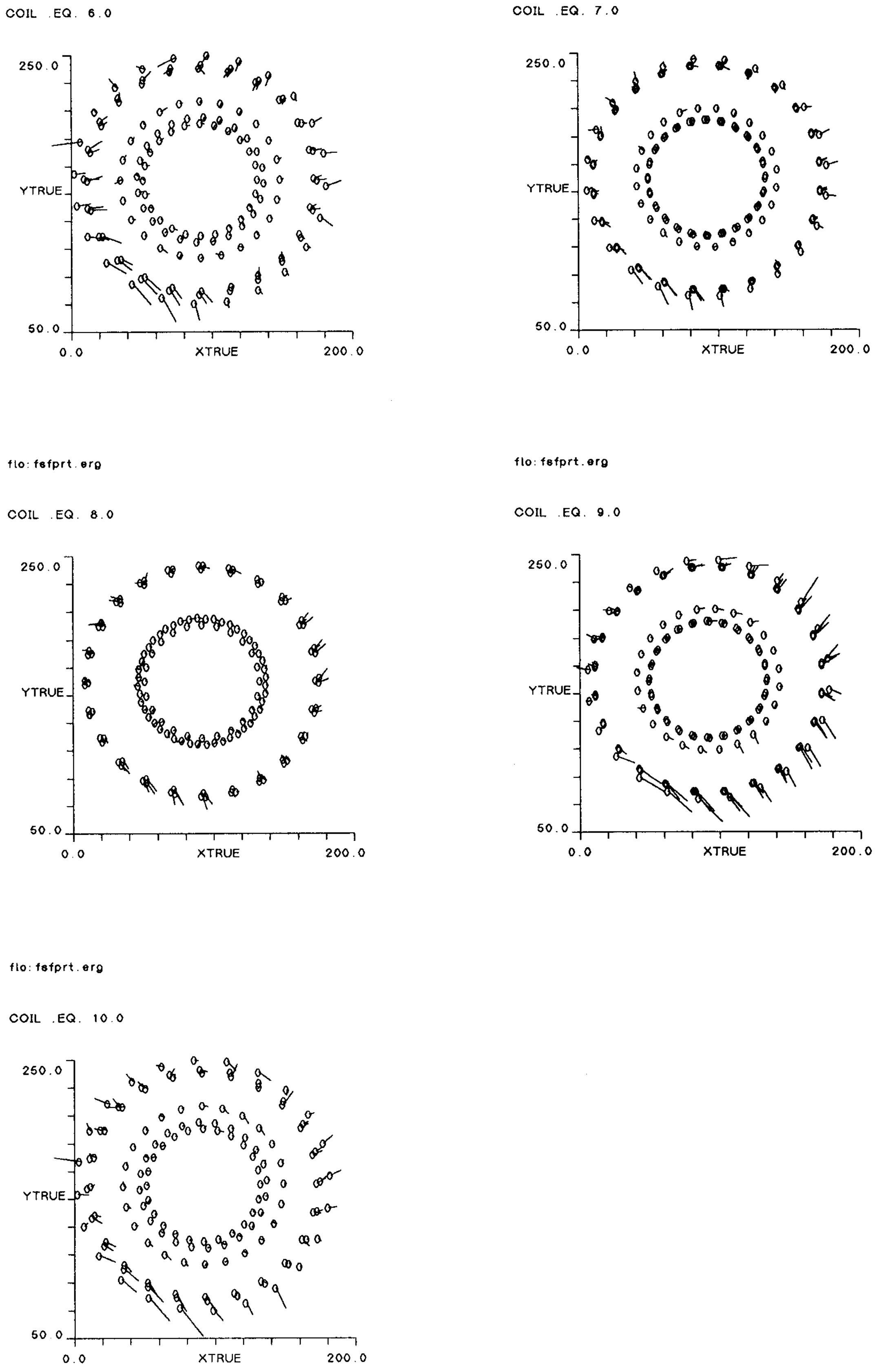

Fig. 4 Error maps for sensors 6 to 10.

The magnitude of the errors in this calibration trial was 0.43mm +/- 0.39 averaged over all measurement positions, but only 0.20mm +/- 0.13 for positions up to 6cm from the centre. For positions more than 6 cm from the centre the average absolute error was 0.67 +/- 0.42.

In practice, following the experimental session, and as part of the post-processing outlined below, the range of positions occupied by each sensor during speech is superimposed graphically on the error map, and the experimenter can choose whether to apply correction factors. In practice, sensors on the tongue hardly ever require correction as they are usually located in the central portion, i.e within about 4cm of the geometric centre, where errors are generally negligible (around 0.25mm)[1].

The articulator most likely to require correction is the lower lip, which can be fairly far out in the lower left quadrant. A choice of two error correction techniques has been implemented. The first one simply involves subtracting the average of the error vectors in the vicinity of the given set of fleshpoint data. The second one uses a regression analysis to give a prediction of x and y error as a function of the x and y position in the region of the articulatory data. This approach will be necessary if the errors do in fact show some systematic change over this region. In an example of this kind, lower lip data recorded in a region of rather high error was corrected (the high errors themselves subsequently turned out to be due to a problem in the construction of one particular helmet, since resolved) using the regression method based on 18 error vectors in the vicinity of the data. The average absolute error in the x-direction declined from 1.349 to 0.352, and in the y-direction from 0.553 to 0.285, so even under rather unfavourable conditions the accuracy in the outer region of the measurement area can be brought close to that obtained in the central region.

In addition to optimizing the accuracy as such, an important benefit of this approach is that potentially less reliable or less accurate sensors or amplifier channels can be identified in advance, and used either for less crucial measurements, or in regions where acceptable accuracy is still attained.

For each experiment, a log file is generated recording the calibration results and recording details of any correction procedures carried out. The principal remaining drawback to the procedures at present (apart from their time-consuming nature) is that the current version of the calibration device does not allow a completely even distribution of measurement locations (refer back to Figs. 3 and 4) - and also not the same distribution for each sensor[2]. (For coils 3 and 8 it is particularly poor - this cannot currently be avoided without an inordinate increase in the time required for the calibration procedures.

---

[1] The helmet attachment developed in Nijmegen helps to ensure that the tongue coils can be located in the centre for any speaker, which as we see, is beneficial for accuracy. In addition, noise in the signal is also noticeably less in the central region.

[2] A more sophisticated, automated device is now available from the manufacturer and has recently been installed at our institute. We will be able to report on experience gained with this new calibration equipment in the near future.

The simple solution is to use these coils for measurements where it is assumed that no corrections will be called for.)

It should also be noted that the current success of the calibration and correction procedure depends on the fact that a given articulator moves only over a fairly small portion of the measurment field. The correction procedures would require substantial modification if movement amplitudes of 10cm or more could realistically occur.

To conclude this overview of our approach to calibration, two further practical problems that were encountered should be briefly mentioned.

First, it proved necessary to take into account the fact that when sensors are placed in the calibration magazine they are usually offset slightly from their nominal true position. This is not surprising as the tolerances in the manufacture of the coil proper and its surrounding casing are not known. Nor, for example, can we assume that two different users will locate the sensors in the magazine in completely identical fashion. The offset involved can be determined by rotating each sensor in turn through a circle with a nominal radius of zero. The amount of movement occuring during this condition indicates the offset of the sensor from the centre of rotation, and thus its offset from its nominal position in the magazine.

The second practical problem we encountered was similar to the one just mentioned in that it also has to do with determining the geometrical properties of the calibration device itself. It emerged that a considerable proportion of the error remaining after taking account of the offset mentioned above was systematically oriented at a tangent to the circles through which the sensors are moved during calibration. Again, we cannot be sure what gives rise to this effect - possibly it is also a question of the tolerances involved in the construction of the helmet or the calibration device. It simply means that the nominal coordinates of the measurement positions defined by the calibration device must be rotated slightly (in effect, a slight rotation of the "clock face" shown in Fig. 2). Fortunately, this rotation angle seens to be a constant characteristic of the combination of helmet and calibration device. So it does not need to be redetermined at each calibration session.

Both the effects mentioned here are quite small (about 0.5 mm for the offset of the sensors in the magazine, and 0.5 to 1 degree for the rotational effect) but taking them into account gives a noticeable improvement in calibration accuracy itself, and also allows the correction procedure outlined above to work more effectively. Also since they have proved to be rather stable over time (sessions), determining them regularly can give useful clues to equipment malfunction or handling errors by the user.

In summary, our procedure is as follows:

1) Perform a preliminary calibration to set up the V-D conversion tables (using 24 positions per sensor on the 4 cm radius circle)
2) Determine the sensor offsets in the magazine using the zero-radius setting in turn for each sensor
3) Recalculate the V-D conversion using the results from (2)
4) Determine the data for the final error maps using 72 positions per sensor on both the 4 and 8 cm radius circles

5) Optionally repeat stages 3 and 4 if the rotational error must be determined or checked

2. Preparation of the subject

The first task in preparing the subject is to have a set of dental impressions made of the upper and lower jaw. The cast of the upper jaw is used to prepare for each subject a customized t-bar or bite-plate, used to determine the occlusal plane in each experimental session. The t-bar is set up so that the subject bites on it with the second molars (on the cross-piece of the T) and with the central incisors (on the stem of the T), two appropriately placed sensors on the T-bar allowing the lower front edge of the upper central incisors to be determined as the origin of the coordinate system used, and the horizonatal axis parallel to the bite-plane (cf. papers by Schönle, Alfonso et al., Perkell et al. in FIPKM 31, 1993 for further references).

The cast of the lower jaw is used when preparing to attach the sensors to the subject: The distance from the front edge of the lower central incisors to the back edge of the lower second molars is measured. With the subject holding the tongue at rest in the mouth, this distance from the incisors is marked on the tongue (using a drop of an oral disinfectant containing a strong violet dye). This corresponds to a tongue position at rest roughly below the junction of hard and soft palate, and provides a good workable anatomical reference. This point is used to locate the third coil from the front (in our standard arrangement with 4 coils on the tongue). The 4th coil is located at least 1 cm posterior to this one. The 1st coil is located 1 cm from the tongue tip, and the 2nd coil equidistant between first and third coil.

The cast of the upper jaw is also used to prepare a mid-line contour of the hard palate, by placing the cast in the measurement plane of the articulograph and moving a sensor coil along the palate. In order to align this trace with the movement data of the subject a measurement is also made with the above-mentioned t-bar in place on the cast.

Subjects new to articulographic experiments were normally given a dry run prior to the recordings in order to familiarize them with the sensation of sensors attached to the tongue and in order to check for unduly active gag reflex or anatomical impediments to the appropriate location of sensors on the gums of upper and lower jaw.

3. Quality Control during the experiment

As frequently discussed elsewhere (e.g Perkell et al., 1993; Hoole, 1993) it is important to monitor articulographic data for possible rotational misalignment of the sensor coils with respect to the transmitter coils. Although a three-transmitter system can correct to a certain extent for measurement errors caused by mislignment, distortion in the measured coordinates will still result if misalignment is severe, particularly when combined with displacement of the sensors from the mid-line. Thus data for which the magnitude of misalignment is large must be treated with caution. In addition, severe *fluctuation* in the amount of misalignment can, for example, indicate that a sensor is no longer securely fixed to the articulator. Accordingly, our software provides the user with an on-line numerical display of the mean and standard deviation of the estimated rotational misalignment of each sensor for each utterance recorded. This information makes it easy to quickly identify potential problems during the course of the experiment; the information is also recorded in a log file to provide a permanent record of this important parameter for future reference (the sample-by-sample misalignment values are also stored as signal streams in the main data files, in addition to the x and y coordinates).

In view of the potentially serious problems caused by misalignment it is important to discourage the subject from making large or sudden head movements. To this end (and for other reasons) we present the speech stimuli to be uttered by the subject one at a time on a computer screen, thus ensuring a constant head orientation during data collection.

4. Post-processing

After conclusion of the experimental session the articulatory data is processed in several ways:

1) The measured movement coordinates can be corrected, if desired, using the error maps set up during the calibration phase (see above). If correction is carried out the estimated accuracy after correction is recorded in a log file generated by the post-processing software.

2) Two sets of coordinate transformations are carried out:

a) The data is rotated and translated so that the origin is formed by the reference coil located on the upper incisors, and the line joining the reference coils on upper incisors and bridge of nose is vertical. This transformation, carried out on a sample by sample basis, corrects for movements of the head relative to the helmet. However, it should be emphasized that only those translational and rotational movements of the head that do not move the sensors out of the mid-sagittal movement plane can be compensated for. One very useful quality-control measure to catch potential problems of this kind is to check that the distance between the two reference coils remains constant (to within 0.5 mm or so). This distance is calculated, displayed to the user (average value for each utterance) and stored for future reference, along with other details of the processing carried out. Fluctuation in this distance would indicate instability in the alignment of the transmitters with respect to the sensors - and thus the possibility of measurement distortion.

b) The movement data is translated and rotated using the data from the occlusal plane measurements (using two sensors mounted on a t-bar, as outlined above) to set the final origin and orientation of the coordinate system

3) The data is low-pass filtered (Fc approx. 60 Hz) and downsampled from 500 to 250 Hz. The reason for using the rather high frequency of 500 Hz in the first place is that the equipment does not include switchable anti-aliasing filters depending on the sample frequency; there is merely a simple low-pass filter with a fixed cut-off frequency of approx. 220 Hz. The possibility of aliasing effects in this kind of equipment needs to be taken seriously in view of the use of high-frequency carrier frequencies and the potentially complicated effects that can occur when possibly imperfectly demodulated signals are combined to calculate the x and y coordinates. The sample frequency of 500 Hz gives the best compromise at present between trying to avoid aliasing effects and retaining real-time operation.

5. Synchronisation of acoustic and articulatory data

During the experiment the acoustic signal is recorded on DAT-tape using high-quality microphone and microphone amplifier. The second channel of the DAT-tape is used to record a synchronisation pulse generated by the PC running the articulatory data acquisition, marking the start and end of each articulograph measurement sequence (typically one speech utterance of a few seconds duration). After conclusion of the experiment the stereo DAT signal is transferred digitally to computer, whereupon the audio channel is downsampled from 48 to 16 kHz and the synch marks are extracted from the second channel. This timing information is used to align the audio data to the articulatory data. This procedure is quite time-consuming in view of the high data-rate of the stereo DAT signal, and in view of the necessity of ensuring that the synch mark extraction has worked properly before processing the data further. However, the results, both in terms of the quality of the audio data, as well as the accuracy of the synchronisation have been very satisfactory.

6. Final processing of articulatory data

The articulatory data delivered by the acquisition program is in multiplexed form, i.e one file contains all channels of x, y and tilt information for one utterance (measurement sequence). In order to make the data available to our analysis and display programs it has to be demultiplexed into signal stream files containing one channel of data (e.g tongue-dorsum x, tongue-tip tilt), each single-channel file containing all the utterances from one experiment.

At this stage the movement data is low-pass filtered once more using a steeper filter than the one used above at the post-processing stage. The spectral distribution of remaining noise in the articulatory signal has varied somewhat from session to session and the data should be inspected at this stage for evidence of low-frequency disturbance (see discussion above of aliasing effects). However, in practice, an FIR filter with Fc of 40 Hz has proved applicable to almost all recordings.

From the position signals velocity signal streams are calculated (x, y and tangential, i.e sqrt (vx**2+vy**2)). These signals are additionally low-pass filtered at 25Hz.

At this point, the basic set of synchronized signal streams is available and the process of segmentation and parameter extraction can start.

II. Data Processing

In this section a typical set of data-processing procedures will be illustrated from a project on German vowel articulation (German Research Council Project "Deutsche Vokale").

1. Acoustic segmentation

A waveform editor is used to mark the location of the target vowel and surrounding consonants in the audio waveform. In addition to providing acoustically defined durational information as counterpart to the kinematically derived durations discussed below, this processing step is used to check the recording systematically for gross mispronunciations and other potentially problematic items e.g voiceless vowels etc. The acoustically defined segmentation is also used to set up a time window within which the kinematic parameters are semi-automatically determined.

2. Kinematic parameters

The basic approach to kinematic parameter extraction is outlined in Hoole et al. (1994) and is summarized briefly here (see Fig.5):

The procedure essentially involves dividing the target CVC units (with C1=C2=p,t,k) into a CV, a VC and a so-called nucleus segment. The latter is simply defined as the segment between the offset of the CV and the onset of the VC movement. CV and VC onsets and offsets are in turn defined using a threshold criterion in the velocity signal of the sensor assumed to be most intimately involved in formation and release of the consonantal constriction (i.e lip, tongue-tip and tongue-back respectively for p, t and k). The threshold criterion used was 20% of peak velocity. The use of this criterion, rather than the more usual zero-velocity criterion encountered in the literature was motivated by the desire to avoid two typical problems in the determination of movement onsets and offsets: firstly vowels with a long steady-state duration may not have an unambiguous choice of zero velocity separating CV and VC movement, but rather an extended period of near-zero velocity, the precise temporal location of zero velocity and hence the measured durations of CV and VC phases thus being very unstable. The criterion used helps to locate the segmentation point at a relatively steep, and thus temporally well-defined point in the velocity curve.

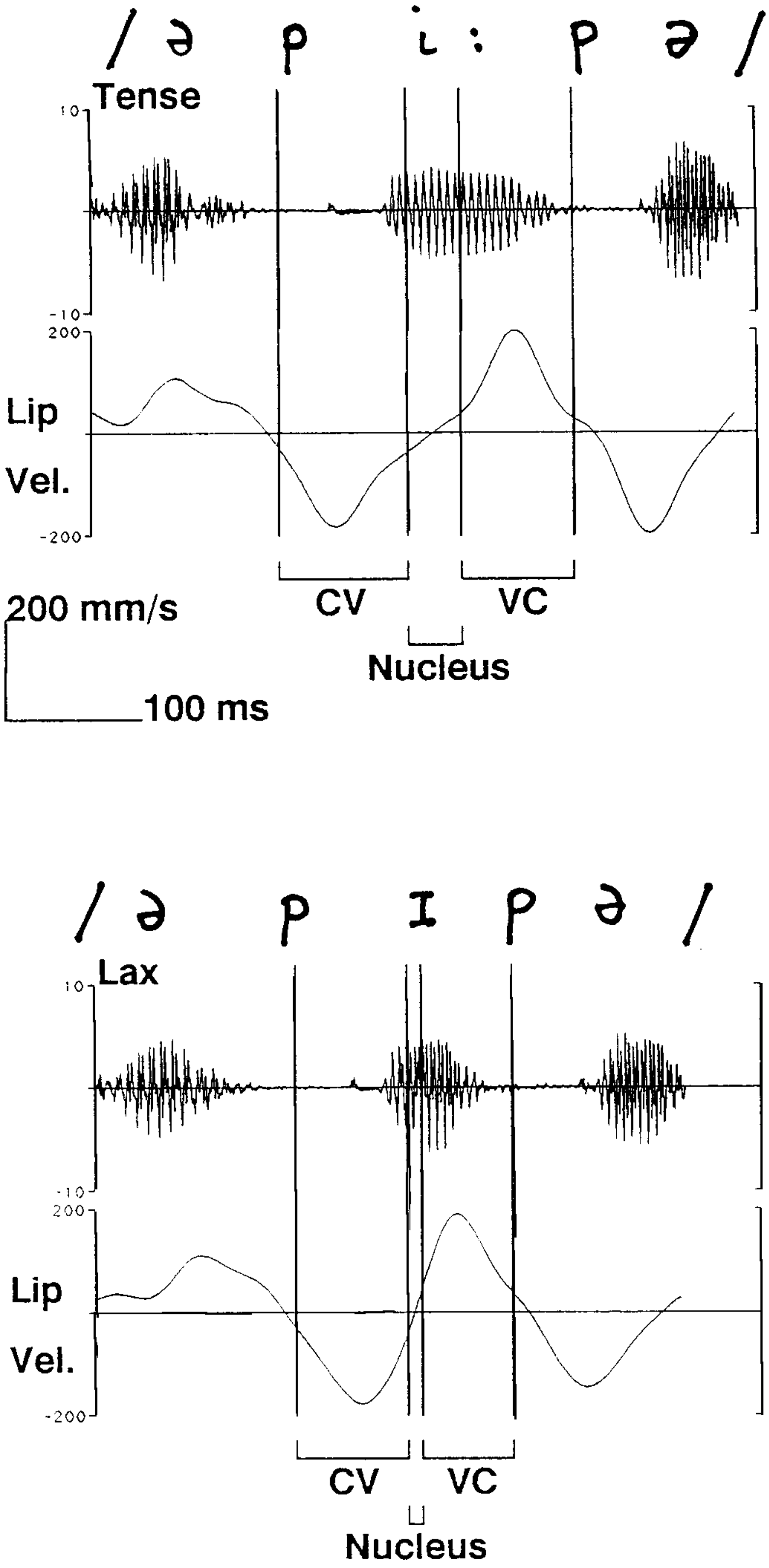

Fig. **5** *Example of division into CV, Nucleus and VC segments. Top: Tense vowel; Bottom: Lax vowel. Speed of lower-lip vertical movement is shown (based on Fig.1 in Hoole et al., 1994).*

The contrary case, sometimes found in short lax vowels, but more often with dorsal consonants (i.e /k/) in the context of back vowels, involves cases where no instant of zero velocity occurs at all - the sound being characterized by continuous movement throughout its duration (see analysis of dorsal consonants in Mooshammer et al., 1995). This necessitates the use of a velocity criterion > 0 as criterion for attainment of a target configuration (in fact, some cases with /k/ still remained intractable to the standard kinematic analysis, not even showing a clear local minimum of velocity during the assumed period of consonantal closure).

Given the kinematic segmentation just outlined, two basic groups of parameters were extracted for each VCV sequence

2.1 Mid-point parameters

The temporal mid-point of the nucleus segment was assumed to represent the point at which the vowel target configuration was attained[3]. The simplest form of analysis thus consists in extracting the coordinates of each sensor at this algorithmically determined time instant, whereupon the data can be passed on to standard statistical packages,or plotting routines for systematically comparing subsets of the data. In addition, the static data forms the basis for the analysis of patterns of variability outlined below.

Although the parameters extracted under the heading of mid-point parameters essentially permit a static analysis of vowel articulation we have also extracted the velocities at the different sensor locations for this time instant, since, as discussed above, it is not always the case that this time instant represents a point of zero velocity. We believe that it should prove an interesting extension to the purely static analysis to examine which vowels, which consonant contexts and which articulators show a greater propensity to non-zero velocity at the time instant assumed to be associated with attainment of the articulatory target. However, this analysis has not been carried out yet.

2.1.1 Variability parameters

These form a category of their own among the mid-point parameters. The main parameter used has been the 2-sigma area of variation as a measure of the magnitude of contextual and token-to-token variability. Its calculation is discussed in the contribution to ICPhS 95 (Hoole & Kühnert, 1995). It essentially involves performing a principal component analysis of the two-dimensional data at each sensor location, and determining the area of the ellipse whose main axis is oriented along the first principal component of variation. This parameter is attractive for having a readily interpretable geometric interpretation in the Cartesian space in which the measurements are made (with our standard

---

[3] It is interesting to note that for short high vowels in the context of consonants with long periods of aspiration, the target configuration was often achieved before voice-onset for the vowel.

choice of ellipse radii corresponding to 2 standard deviations along the two components of variation the ellipses can be expected to enclose approximately 90% of the data, given a two-dimensional normal distribution). In addition we have made provision for future analyses, in which the articulatory and acoustic variability are to be related, to supplement the simple parameter of ellipse area by two further parameters readily derivable from the principal component analysis: these are the orientation of the ellipse in Cartesian space, together with the shape index of each ellipse, defined as the length ratio of major and minor axes. These parameters can help to quantify the extent to which major variation is parallel to the vocal tract wall (especially the hard palate) as well as the extent to which the dimension of main articulatory variation may represent a dimension of low acoustic variation.

2.2 Movement parameters

The parameters derived under this heading are oriented towards capturing the salient characteristics of the complete CVC movement pattern (cf. Hoole et al., 1994)

The parameters are:

a) Duration of the CV, nucleus and VC segments
b) Displacement and peak velocity of the CV and VC segments
c) Parameters derived from the above elementary parameters:
    i.e the ratio of peak velocity to displacement ("stiffness")
    the ratio of peak to average velocity (referred to as "**c**" in the ICSLP paper). The latter should potentially give a means of assessing the underlying similarity of movements differing in absolute values of duration, peak velocity and displacement. However, this parameter does not capture all possible relevant differences in velocity profile. Thus further work may need to make use of additional parameters e.g measures of the symmetry of acceleration and deceleration phases of a movement. However such approaches may turn out to be very difficult to apply to all the data consistently as the difficulties already discussed above with the respect to the segmentation of velocity signals will tend to become exacerbated as higher-order derivatives with respect to time are considered.

References

Alfonso, P., et al. (1993) "Calibration, validation and hardware-software modifications ot the Carstens EMMA system", FIPKM, 31, 105-120.

Hoole, P. (1993). Methodological considerations in the use of electromagnetic articulography in phonetic research. FIPKM 31: 43-64.

Hoole, P., Mooshammer, C. & Tillmann, H.G. (1994) Kinematic analysis of vowel production in German. Proc. ICSLP 94, Yokohama, 1:53-56.

Hoole, P. & Kühnert, B. (1995) "Patterns of lingual variability in German vowel production", Proceedings XIIIth Int. Conf. Phon. Sci., Stockholm, 2, 442-445.

Mooshammer, C., Hoole, P. & Kühnert, B. (1995). *On loops*. J. Phonetics, 23: 3-21.

Perkell, J., et al. (1993) "On the use of electrical midsagittal articulometer systems", FIPKM, 31, 29-42.
Schönle, P. (1993) "The developmental genealogy of electromagnetic articulography (EMA)", FIPKM, 31, 83-90.