Chapter 20

Five-dimensional articulography

Phil Hoole and Andreas Zierdt

Abstract

Current developments in the use of five-dimensional electromagnetic articulography for speech research are reviewed. Obvious advantages are the higher information density per sensor (three Cartesian coordinates, two spherical coordinates) compared to traditional 2D EMMA systems, and removal of the necessity to constrain the subject's head. The drawbacks are equally related to this higher dimensional space: position calculation involves solving a non-linear optimization problem. In some cases, unstable solutions are encountered, resulting in mistrackings.

On the positive side, we illustrate how the higher information density allows particularly succinct and robust characterizations of tongue configuration. Discussion also focuses on monitoring of head movement. This is crucial for accurate recovery of articulator movements themselves, but is also intrinsically interesting as part of speech motor activity. In addition to improving the naturalness of the speaking situation, the freedom of head movement also means that subjects tolerate longer recording sessions. This can facilitate new experimental paradigms.

Regarding drawbacks (and ways around them), instabilities in position calculation are illustrated and it is shown how a first estimate of the measured positions can be used as a starting point for a more robust estimate, taking the continuity of speech movements into account. Diagnostics for assessing the reliability of the final solution are outlined. While work remains to be done to ensure the same accuracy over the whole 5D-measurement space, it is concluded that the system already offers unparalleled scope for large-scale acquisition of flesh-point data.

20.1. Introduction

Electromagnetic midsagittal articulography (EMMA) systems have now been in routine use for phonetic research for some 15 years. With the increasing availability of so-called three-dimensional (3D) systems (in the form of the commercially available AG500 system [Carstens Medizinelektronik] and the Aurora system [NDI]) it seems now an opportune time to take stock of some of the specific advantages of this latest generation system, but also of the additional complexity in processing the data. Both of these aspects are encapsulated in the fact that a system such as the AG500 is better regarded as a five-dimensional (5D) system. On the one hand, the amount

AQ: Are we to retain bold?

Yes please

AQ: Please confirm insertion.

OK

of information provided by each sensor is substantially increased compared to the old 2D systems, consisting now of three Cartesian coordinates (x, y, and z) and two angular coordinates (azimuth and elevation). Within the overall context of flesh-point measurement systems, this gives such an electromagnetic system some very useful advantages, practical examples of which will be presented in the first section below. On the other hand, deriving this 5D information for each sensor from the raw data (consisting of the signal induced by six transmitter coils) consists of a search for an optimal solution in a correspondingly high-dimensional space. As Kaburagi et al. (2005, p.440) have remarked, 'a complicated non-linear problem must be solved'. Some of the difficulties that typically need to be grappled with are illustrated in the second part of the chapter.

20.2. **Benefits of high-dimensional sensor data**

In this section, we will first recap by way of background the main differences between the new 5D systems and old 2D systems. Secondly, we will illustrate the use of the angular coordinates that are new to the 5D system from two points of view: (1) capturing phonetically relevant aspects of tongue shape; (2) potentially increasing the robustness of the data. Thirdly, we will discuss issues related to the monitoring of head movement. The fact that the higher dimensionality of the data means that the subject's head no longer has to be constrained is perhaps the simplest but nonetheless quite possibly the most useful advantage of moving beyond measurements in two dimensions.

20.2.1. **Background to EMA and EMMA**

In a traditional two-dimensional (2D) EMMA system the movements of sensors attached to structures on the midsagittal plane of the subject (e.g., upper and lower lip, jaw, three or four locations on the tongue) can be tracked at sampling rates of typically at least 200Hz.[1] For best accuracy, the main axis of the sensor coils must remain aligned in parallel with the main axis of the transmitter coils. In other words, the midsagittal plane of the subject must be kept in alignment with the measurement plane of the system. This makes it necessary to have a firm attachment between the head of the subject and the helmet holding the three transmitters. The three-transmitter design makes it possible to compensate to a certain extent for movements off the measurement plane, and for misalignment of the sensors, but such changes in lateral position or in alignment cannot actually be directly measured (see e.g., Hoole and Nguyen [1999] for more background).

In the newer 5D systems exactly the same sensors can be used, but by using more transmitters, namely six, it is possible to recover the theoretical maximum information yield of such a sensor. This consists of the three positional coordinates (x, y, z) and two angular coordinates. The latter coordinates in effect define in a spherical coordinate system the direction in which the sensor is pointing. The coordinates are often referred to as azimuth, which is angular position in the xy-plane (normally the axial or transversal plane of a human subject), and elevation, which is angular displacement out of the xy-plane. For the kind of sensors currently in use in the AG500 system (and previous 2D EMMA systems) there is a further rotational degree of freedom of the sensors which cannot be captured: The system is 'blind' to rotations about the main axis of the sensor, since this does not cause any change in the electromagnetic induction.

In a 2D EMMA system, there is essentially no choice in terms of orientation when attaching sensors to the tongue: As just mentioned, the main axis of the sensor must be parallel to the

[1] The sampling rate for the AG500 system is currently fixed at 200Hz, but it may be possible to increase it in future. The sampling rate of the Aurora system was initially much lower (see Kröger et al., 2008), but has increased in more recent revisions.

transmitter coils. In a 5D EMA system there are different possibilities: If the main axis of the sensor is aligned parallel to the midline of the tongue (essentially at right-angles to the orientation used in the 2D system), then the elevation component of the angular coordinates in effect represents the angle of a tangent to the midline contour of the tongue at the sensor location, ~~and thus the angular component is~~ most closely related to the midline shape of the tongue. This information will be used for the detailed example in Section 2.2 below.

AQ: Please check renumbering. This has been done as we changed the numbering of Introduction from 0 to 1.

OK, but wouldn't 20.2.2 be better?

20.2.2. Use of angular coordinates

In this section, we illustrate the use of the angular coordinates derived for each sensor to provide phonetically useful information, since this new feature of the system has not yet found wide usage in analysis of speech movements. By way of example, we will use some data on consonantal articulation in Moroccan Arabic (see e.g., Zeroual et al., 2007). In Arabic, the so-called emphatic consonants form an interesting feature of the consonantal system. The emphatic consonants are traditionally regarded as being distinguished from the non-emphatic counterparts by pharyngealization, i.e., a secondary articulation involving retraction of the tongue root. However, it has been suggested that there are also characteristic differences at the front part of the tongue, with the emphatic (coronal) consonants being apical and the non-emphatic counterparts laminal.

Figure 20.1 shows data for ten consonant categories as spoken by one speaker (seven repetitions per category, all spoken intervocalically between low vowels). The labels used in the figures are as follows: lower-case 't, d, s' label the non-emphatic consonants and upper-case 'T, D, S' their emphatic counterparts. In addition a nasal, a lateral, a flap and a postalveolar fricative are shown, labelled 'n, l, r, sh', respectively.

The top left panel of the figure shows a traditional mid-sagittal view of the tongue-tip. Mostly there is quite a clear separation of the categories, but there is some overlap between non-emphatic /s/ and emphatic /S/, and emphatic /T/ overlaps not only emphatic /D/ but also non-emphatic /d/.

In the top right panel of Fig. 20.1 we have plotted the elevation component of the angular coordinates for two sensors: that from the tongue-tip on the y-axis, that from the tongue-back (the rearmost tongue sensor in this experiment) on the x-axis. As an aid to interpretation the four corners of the schematic plot at the bottom right of the figure show the orientation of these two tongue sensors at the extreme values of the x and y axes, corresponding to the orientation of the sensors that would be seen by an observer viewing tongue configuration in the sagittal plane (as outlined in the previous section, these orientation vectors here represent a tangent to the midline contour of the tongue at the sensor locations). Thus low values on the y-axis indicate that the tongue-tip is oriented more or less horizontally, while high values indicate that the tongue-tip is angled up, presumably corresponding to a more apical configuration of the tongue. It is interesting to note that the separation between the consonant categories is even more clear-cut in this representation than in the previous Cartesian representation of the tongue-tip: for example, the distinction between emphatic /S/ and non-emphatic /s/ is clearer, and also between non-emphatic /d/ and emphatic /T/, as well as between /n/ and other consonants. Thus, these two angular parameters appear to give a very succinct characterization of important features of the tongue shape.

Many variations on this general approach are conceivable, although they cannot be presented here: For example, if sensors are mounted at right-angles to the midline of the tongue and not only on the midline itself but also lateral to it, then angular information could help to reconstruct the tongue shape in the coronal plane.

A further attractive feature of the angular coordinates is that – at least in some experimental situations – they represent very robust information. Let us imagine a worst-case scenario in which all reference sensors used to compensate for head-movement failed. This situation is simulated in the bottom left panel of Fig. 20.1 for the traditional sagittal Cartesian coordinates of the tongue-tip.

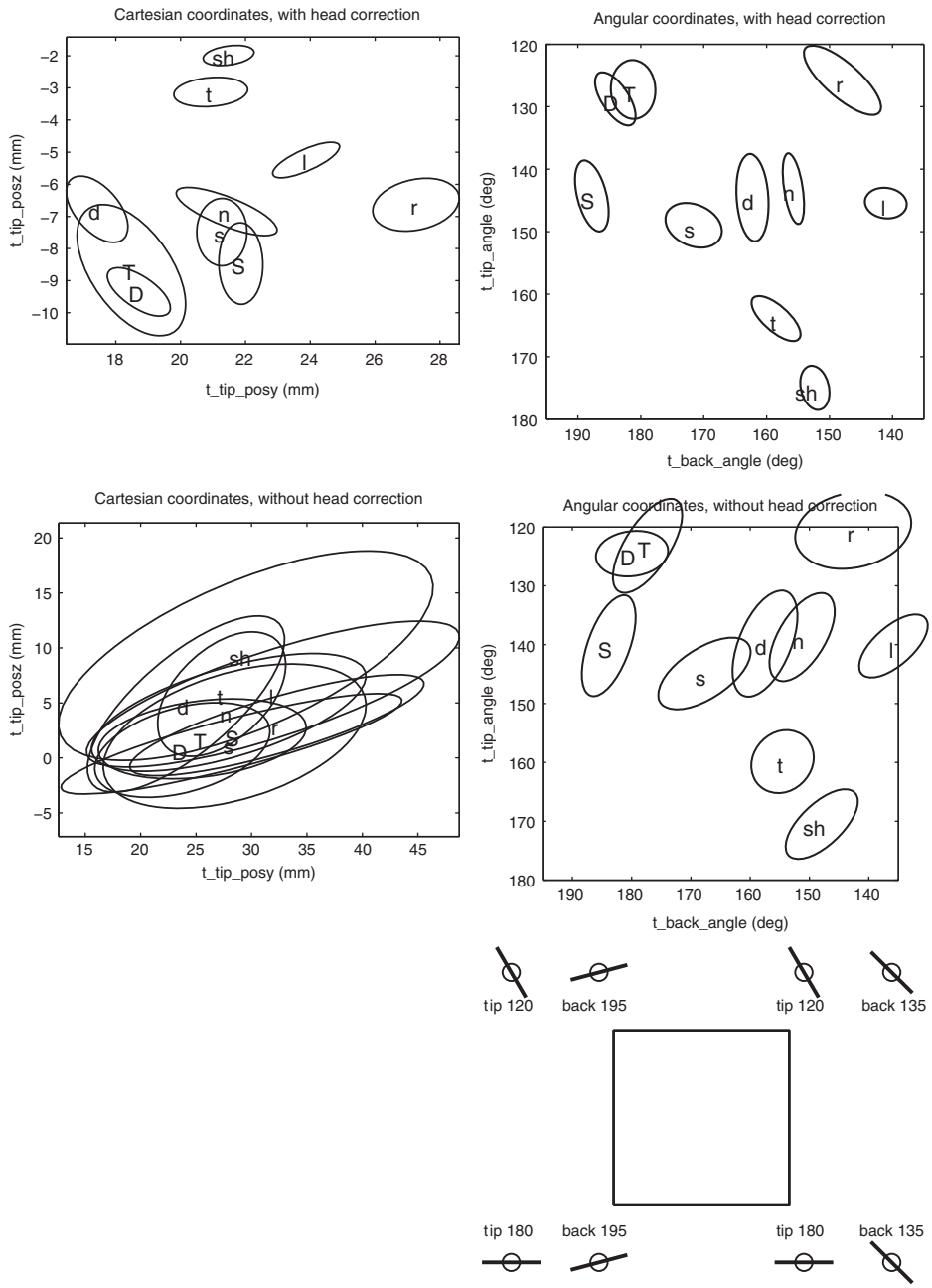

Fig. 20.1 Left column: Traditional sagittal view of position of tongue-tip for articulation of Moroccan Arabic consonants in /aCa/ context. See text for explanation of category labels. Anterior is to the left. Right column: Representation of tongue configuration using orientation of the tip and back sensors in the mid-sagittal plane. Top row: Tongue data after compensation for head movement. Second row: Data without compensation for head movement. Bottom right panel: Schematic illustration of tongue sensor orientations in the sagittal plane, corresponding to the extreme positions on the x and y axes in the two panels above.

Clearly, any regularities in the data have essentially disappeared. This is not surprising: even in a simple lab-speech experiment with the subject sitting quietly in a chair reading prompts from a computer screen then a few centimetres of head translation is likely to occur over the course of the session. This will essentially swamp any linguistic distinctions, which – as can be seen in the top left panel – will normally play out over a range of no more than 1 cm.

On the other hand, in this kind of setup changes in the orientation of the head are likely to be quite small, or at least small relative to the linguistically induced range of orientations which here amounts to about 50 degrees for both sensors. Accordingly, the bottom right panel of the figure shows the same angular data as in the top right panel, but without head-movement correction. It will readily be observed that the separation between the consonant categories is still very clear-cut. While many subjects may change the orientation of their head more than this subject (who rarely exceeded 5°, as indicated by the fact that it was possible to keep the same axis scaling for top and bottom panels on the right), it can still be expected that the robustness of the orientation information will exceed that of the translational information, if thought of as the ratio of linguistically induced changes to posturally induced changes.

This has admittedly been chosen as an extreme example; nevertheless, the next section will show that it is certainly realistic to expect – if not complete failure – then at least some variability in the quality of head-movement correction.

20.2.3. **Monitoring head movement**

In this subsection, we will use monitoring of head movement as a topic to firstly drive home the point about the high information density per sensor, going on to point out some practical issues in correcting for head movement. We will then give an example of where head movement can represent linguistically relevant behaviour in its own right. Finally, we will briefly point out interesting experimental paradigms that become much more feasible when the subject no longer needs to be physically attached to the transmitter assembly.

20.2.3.1. Capturing the degrees of freedom of head movement

In a 2D EMMA system, compensating for head movement involves capturing translation and rotation in the sagittal plane, i.e., two translations and one rotation, typically using sensors mounted on the upper incisors and the bridge of the nose. In the new system, head movement is not restricted, so the requirement is to capture the full six degrees of freedom that characterize rigid-body motion, consisting of three translations and three rotations. In analogy to a ship or aircraft the three rotational components of the head considered as a rigid body can be defined as,

(1) pitch: rotation about an axis perpendicular to the sagittal plane, i.e., about an axis parallel to the lateral axis (the main component of nodding movements to indicate 'yes')

(2) roll: rotation about an axis perpendicular to the coronal plane, i.e., about an axis parallel to the anterior-posterior axis (inclining the head towards the shoulder)

(3) yaw: rotation about an axis perpendicular to the transversal plane, i.e., about an axis parallel to the vertical axis (the main component of shaking movements to indicate 'no').

It is instructive to consider how many sensors are necessary to capture all six degrees of freedom. One might assume that because more degrees of freedom need to be captured than in a midsagittal EMMA system then more sensors must also be required. In fact, this is not the case (though an important caveat will be introduced shortly): Two sensors are in principle sufficient to completely capture head movement; but this depends crucially on the additional angular information provided by each sensor. Consider the typical setup in a 2D EMMA system mentioned above, i.e., with reference sensors attached to upper incisors and nose. Clearly, if each sensor only provided

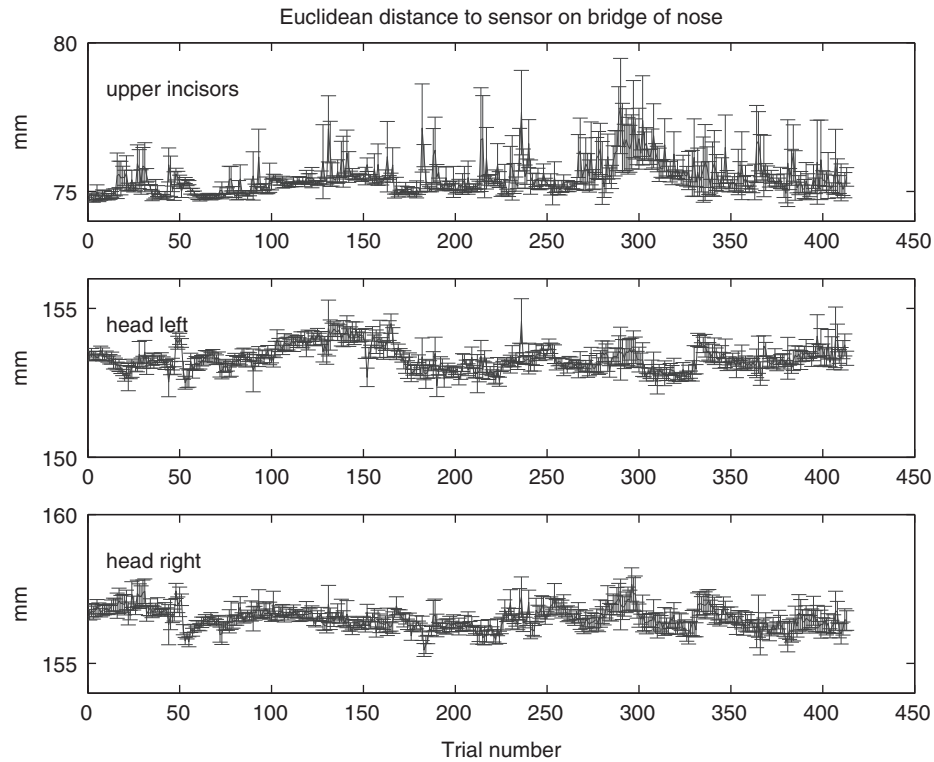

Fig. 20.2 Euclidean distances between pairs of reference sensors.

Cartesian coordinates (x, y, z) then some kinds of head movement would not be recoverable with only two sensors, in particular if the head were to rotate about an axis parallel to the line joining the two sensors. In terms of the definitions given above, this would roughly correspond to yaw.[2]

However, if the main axis of the sensors is mounted at right angles to the line joining the two sensors then the missing information on yaw will be captured in the changes in the angular coordinates of the sensors.

In practice, just using this minimum of two sensors is probably not advisable. Quite apart from the possibility of sensors failing or becoming detached during the experiment it has emerged from experience with many experiments where we typically use four reference sensors (usually sensors mounted just behind each ear in addition to the two traditional EMMA ones) that often the quality of the coordinate data from the different sensors is not equally good. This can be found, for example, from examining the Euclidean distances between all pairs of reference sensors (the distance between any pair of sensors should of course be constant as they are regarded as being

[2] Another way of looking at this is as follows: the total number of coordinates for two purely Cartesian 3D sensors is 6, but when attached to a rigid body this is insufficient to recover the 6° of freedom, since the distance between the sensors is fixed. Thus, in effect one degree of freedom of the two sensors is lost. 5D sensors with 3 Cartesian and 2 angular coordinates provide enough information to make up for this loss. Additional note: For the NDI Aurora system more complex six-degrees-of-freedom sensors are available for special purposes, such as head movement correction. Of course, a lot then depends on this sensor operating reliably.

AQ: Please review change.

Please cancel change. 'Degree' is not a unit of measurement here

AQ: Should degrees be changed to symbol as above?

No!

attached to a rigid body). To illustrate this, Fig. 20.2 shows the mean and standard deviation per trial of the distance between a reference sensor on the bridge of the nose and three other reference sensors over the course of a session consisting of about 400 trials. Clearly, the upper incisor sensor is much less stable than the other two. However, since more than the bare minimum number of reference sensors has been recorded in this case then one could afford to leave out the least reliable one from the head-correction procedure.

20.2.3.2. Relations between head movement and linguistic behaviour: a topic in its own right

Head movement can, of course, itself form part of communicative behaviour and be worth examining in its own right, rather than just being regarded as extraneous movement that needs to be factored out of the measurement of articulatory movement (see e.g., Munhall et al., 2004).

In this section, we use material from a recent study of the articulatory correlates of tone production in Mandarin Chinese as an example for this kind of investigation (see Hoole and Hu 2004). The main thrust of this investigation was to look for different patterns of tongue movement related to the four tones of Mandarin. For this, data was acquired with the AG500. Since head movement had to be monitored anyway the opportunity was taken to examine this as well for any evidence of regularities related to tone production because there is some evidence in the literature that limited tonal perception is possible based on visual information alone (e.g., Burnham et al., 2001; see also Mixdorff and Charnvivit 2004). This raises the question of what visual information actually could be involved here, and head movement is certainly a candidate (see e.g., Munhall et al., 2004 and Yehia et al., 2002) for some evidence that F0 can be quite well predicted from head-motion parameters – at least on an utterance-by-utterance basis).

In a corpus of isolated words spoken by one speaker we found a consistent difference in head position for tone 3 (the low dipping tone) compared to the other tones, particularly compared to tone 2 (the high rising tone): the position of the head tended to be lower and more posterior for tone 3. This is illustrated in Fig. 20.3.

Of course, it is scarcely reasonable to expect that speakers adopt a specific *position* of the head for the different tones (as extracted here at the midpoint of the vowel), or to expect that head movement somehow mimics F0 movement. Thus, it is perhaps of more interest that the overall pattern of movement tended to separate tone 3 from the other tones: Velocity of head movement in the forward and upward direction tended to be higher for tone 3. Velocity patterns are attractive in the context of visible behaviour of this kind because they may be communicatively more robust than subtle differences in position (cf. Keating et al., 2003). Currently we do not know whether these results will generalize to other speakers and other speaking styles. But we find it perfectly plausible that precisely when a speaker is aiming at a very clear style of speech then characteristic patterns of visible behaviour may be coherently integrated with articulatory and phonatory behaviour (see also Craig et al., 2008). And the simple methodological point to be made here is that even these tentative results show that with this kind of measurement system a widening of the perspective on speech-related motor behaviour is worthwhile.

20.2.3.3. Potential for new experimental paradigms

Freedom of head movement, by increasing the subject's comfort, quite simply means that longer experiments are possible. We have successfully run sessions in which over 1,000 utterances were recorded, and in which the net speaking time (i.e., leaving out pauses between utterances) amounts to about 50 minutes. Clearly, this can provide a more solid statistical foundation for almost any kind of experiment. In addition, it makes paradigms much more feasible in which there may be a rather low ratio of 'interesting' material to total recorded material, for example,

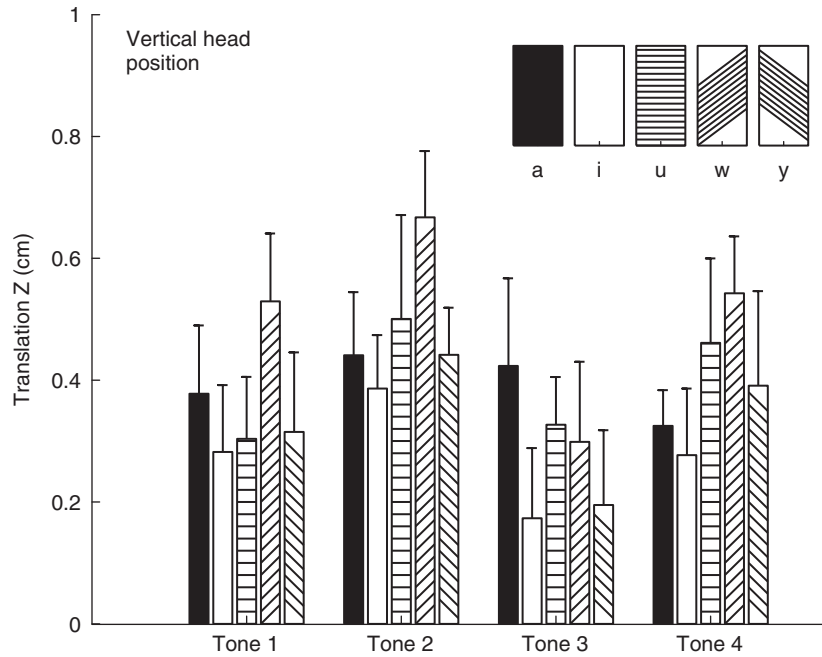

Fig. 20.3 Mean vertical position (in cm) of head at midpoint of vowel ('w' symbolizes a back unrounded vowel) for the four tones of Mandarin (12 repetitions). Before averaging, data were normalized by subtracting the mean of each block of repetitions. Error bars indicate standard deviation. Differences between Tone 3 and Tone 2 significant at $p < 0.01$.

in experiments to elicit speech errors, and in recordings based on spontaneous speech. A further area that could be mentioned here is that of speech technology: the amount of data that can now be acquired is starting to get within the ballpark of what is typically required of training data in applications of this kind (e.g., Ling et al., 2008). Regarding spontaneous speech paradigms, being recorded with a device such as the AG500 will, of course, always be very much a laboratory situation for the subject. Nevertheless, freedom of head movement is certainly an important contributor to utterance naturalness in unscripted speech situations (we have, for example, started to work with a map-task [Anderson et al., 1991], in which one dialogue partner gives directions for a route to follow to the other partner).

Given that an EMA system represents a considerable investment for most labs it appears worthwhile to be alert to possibilities for expanding the range of applications of such a system and we conclude this section with one brief example of further directions that diversification could take.

The example shows that sensors (with a certain amount of circumspection) can be quite easily attached to the eyelids, allowing the time instants of blinks to be detected. In the example in Fig. 20.4 the trace labelled 'T5 raw' is the signal from the eyelid sensor time-aligned with the audio trace. The sharp peaks represent blinks occurring about 0.5 s before the onset of the utterance and roughly at the location of the stressed first syllable in 'relatively' at about 1.75s on the time axis. For an application of this kind, where one may just be interested in the temporal location of a sharply defined event it is probably not even necessary to calculate the sensor positions. The trace

AQ: Please review change.

OK

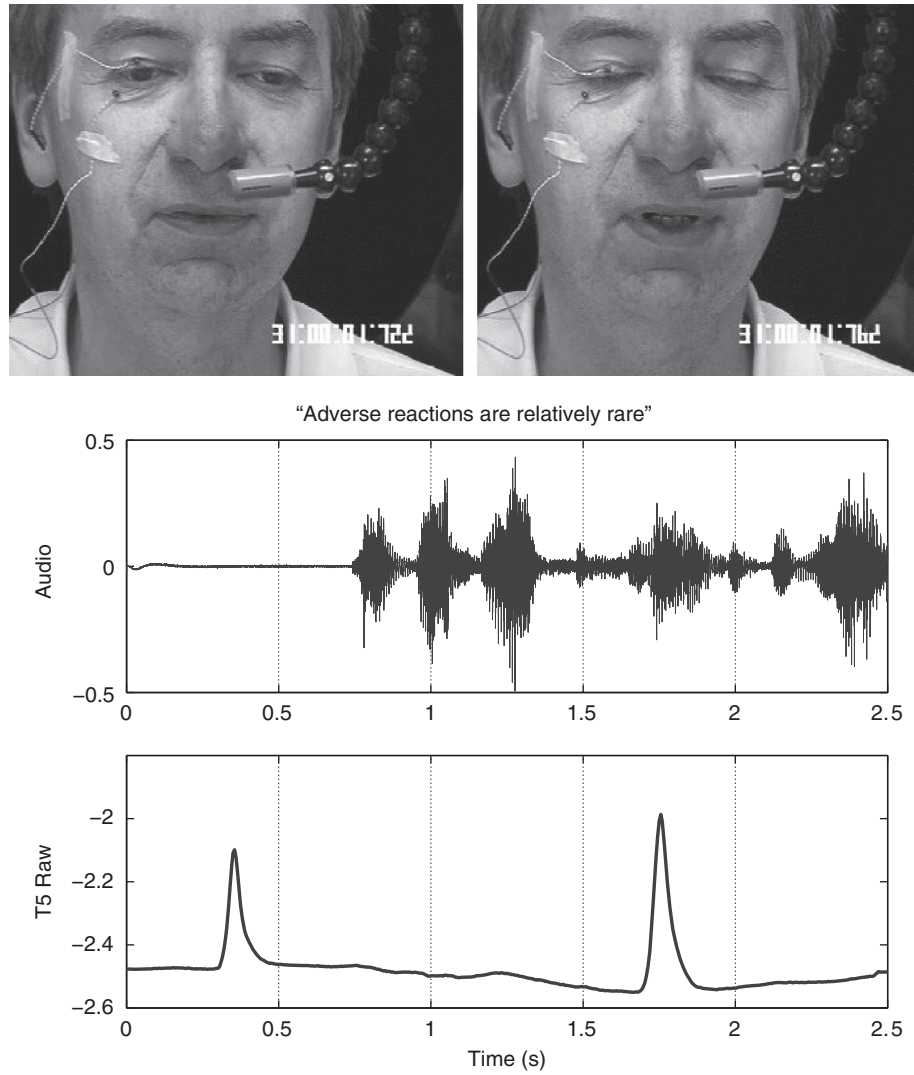

Fig. 20.4 Using an EMA sensor to capture blinking movements. Two consecutive frames from synchronized video. The second frame is located at about the peak of the second blink event at around 1.75s.

used in the figure simply shows the raw signal induced in the sensor, chosen with respect to the transmitter that gave the strongest modulation of the signal for blink events. Another point in this figure that may be of methodological interest is that our experimental setup allows us to synchronize video filming (at 25 frames or 50 fields per second) with the EMA data acquisition. The last four digits in the bottom right corner of the two video stills are generated by a video timer (FORA VTG53) that is triggered by the synchronization signal from the AG500 hardware to start counting at the onset of each EMA trial. They thus indicate the location in seconds of the images with respect to the time traces in the two panels below.

20.3. **Position calculation as a problem in non-linear optimization**[3]

In a 2D EMMA system, calculation of the sensor positions (x and y) just requires applying a simple geometrically based formula to the raw input signals from the three transmitters. The down side to the high dimensionality of the information provided by each sensor in the new system is that a closed-form solution deriving the three sensor positions and two sensor orientations directly from the raw input of the six transmitters is not available, but rather requires a search in a correspondingly high-dimensional space. Essentially, one is searching (for each sensor, at each time instant) for the set of positions and orientations that when plugged into the equation for the electromagnetic field will generate the actually measured signal from each of the six transmitters to within some appropriate tolerance criterion. In this sense, position calculation is a member of the large category of inverse problems, i.e., model parameters must be estimated from the observed data. Because of the presence of non-linear terms in the field equation, we are faced with a non-linear optimization problem. In this section we will first of all use a 'toy' example to give a general appreciation of the kinds of difficulties that can arise in such procedures, and then go on to consider two specific problem areas that have been encountered quite frequently in practical work with large amounts of EMA data and show some potential solutions.

20.3.1. **Non-linear optimization: a toy example to illustrate the basic scenario**

For the example in this section, we will use the following conventions:

x Position

$f(x)$ A non-linear function of x (e.g., the magnetic field model)

y Measured sensor signal

To determine x, solve $f(x) - y = 0$

The reason for referring to it as a 'toy' problem is that for illustrative purposes the solution search evolves in one dimension, whereas in practice we have a search space of much higher dimensionality to contend with (also neither of the non-linear functions used here as examples are identical to the magnetic field function, but they are realistic enough to illustrate important features).

The basic procedure is illustrated in the left panel of Fig. 20.5. This procedure for finding a solution is based on Newton's method. In practice, we currently generally use the Levenberg-Marquardt approach implemented in Matlab's optimization toolbox since this has generally proven to be more robust, but this makes no difference to the illustration of some basic concepts and difficulties. The search begins at some start value for x (labelled x_0 in the figure). Here $x = 0$ is used, but this is simply the choice that may be made if one has no prior information to go on. As will become apparent below, the choice of start position can be crucial. At the start position, the algorithm makes an estimate of the zero-crossing of the function using the gradient of the function at the start position. The corresponding x location (x_1) is then used to re-estimate the zero-crossing and the procedure iterates until some termination criterion is reached.[4] The right

[3] Discussion in this section is based on our experience with the Carstens AG500 system. The situation with the NDI Aurora system is substantially different since, as far as we are aware, the user has access only to the final sensor coordinate values, and not to the raw signals. It seems that the system sometimes outputs missing data, presumably, when the result of the position calculation is considered unreliable (see Kröger et al., 2008). Recovery of missing data by means of additional off-line processing does not appear to be possible.

[4] The termination criterion may be just based on the number of iterations, but typically also involves some threshold for improvement in the residual from one iteration to the next (or of change in the calculated position).

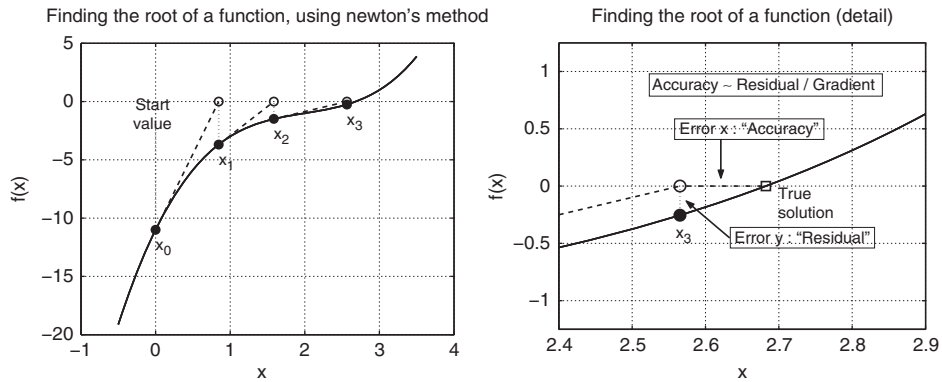

Fig. 20.5 Left panel: Illustration of a few iterations towards a solution using the Newton procedure. Right panel: Zoomed view of the true solution and the calculated solution.

part of the figure zooms in on the situation at termination which is assumed to occur here at x_3. The accuracy of the solution (e.g., positional accuracy in mm) can then be defined as the horizontal distance along the x-axis between the x-value at termination and the true solution at the function zero-crossing. Although accuracy is intuitively the measure one is interested in it can in practice never be known, except in special calibration setups. What is known is the measure referred to here as the residual, i.e., the value of the function f(x) at the termination value of x (i.e., the amount by which the function deviates from the ideal value of zero). In the specific case of the EMA system the position and orientation calculation for each sensor results in a residual with respect to the signal from each of the six transmitters. The root mean square over these six values is what the optimization procedure aims to minimize. This is by no means the only way of formulating the optimization problem (e.g., one could imagine minimizing the worst residual) but none of the alternatives examined to date have proved obviously superior. Another important point illustrated in the right panel of Fig. 20.5 is the relationship between accuracy and the residual: this depends on the *gradient* of the function in the vicinity of the solution. Thus, different parts of a recording where the position calculation results in similar RMS residuals may not be equally accurate. In particular, even low RMS values can have relatively high inaccuracy if the gradient at the solution is shallow. This needs to be kept in mind, since the RMS value is the principal quality criterion available to the user.

AQ: Please provide expansion of RMS.

Insert 'root mean square' in brackets

The previous illustration illustrates the ideal case where the algorithm will converge quickly on a good approximation to the true solution. A more difficult situation is illustrated in the next example (Fig. 20.6). Here, the iteration step starting at x3 fails to get closer to the true solution; rather, x_4 ends up back close to x_2. The algorithm may then become stuck in this region. The right panel of the figure again zooms in on the true and calculated solution. This illustrates from a different perspective the fact that the value of the residual can be quite small, but may nevertheless be associated with poor accuracy. This example also makes clear that the choice of start position can have a major effect on the result. For any given true solution, it is possible to define a convergence region, i.e., the region within which the algorithm is bound to converge on the solution. For start positions outside the convergence region it can be a matter of chance whether one of the iteration steps manages to jump into the convergence region, or whether the iterations become locked into a highly erroneous region where the algorithm then terminates because the residual fails to improve noticeably from one iteration to the next.

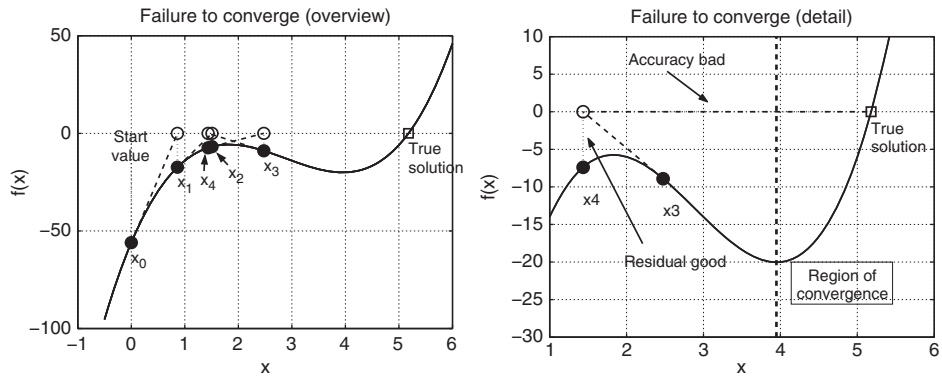

Fig. 20.6 Left panel: Example of failure to converge on true solution. Right panel: Zoomed view of the true and erroneous solution.

What strategies can be followed to choose the start position?

AQ: Is this a heading? Should it be numbered?

No, it was intended as a single line paragraph. May be better to combine with following paragraph (starting 'First of all')

First of all, for bootstrapping the procedure we have found the following approach useful: Within each trial for which positions are to be calculated it is possible to determine a sample that is likely to give a robust solution by examining the gradients of the function. Calculation can then proceed forward and backward in time from this location. In addition, in many speech experiments it is possible to have a rough idea in advance of the positions and orientations of the sensors. In particular, a ballpark idea of the orientations appears to be very advantageous and is actually easier to specify than the positions, since they are less dependent on where the subject happens to sit within the transmitter assembly. This start-position (which will normally only be used for the first calculated sample in a trial) can then be refined by calculating a few trials from a whole experiment, and – assuming a reasonable proportion of the data looks reliable – then using the average position from these trials as start position for the bulk of the trials. Once position calculation for a particular trial has been initiated then the default procedure is to use the position of the sample just calculated as the start position for calculation of the next sample. Clearly, this will in most cases give a start position that is fairly close to the solution, but in the light of the above remarks it should be immediately apparent that there are two potential pit-falls: Firstly, if the articulator is moving fast and the convergence region is small, then the position of ~~samplen~~ may be outside the convergence region of sample n+1; secondly, if an erroneous position has been latched onto, then the solution may became stuck in this region for some period of time, particularly if the sensor is not moving very much. There may then be a sudden jump in the data when the raw signals have changed sufficiently for the true solution to be located in a more robust region of the search space.

AQ: Please review change.

OK, but please insert space in 'sample n' at beginning of line

There are various additional strategies with respect to the start-positions for getting round these difficulties,[5] but none are foolproof, and since current versions of the optimization procedures do not incorporate any continuity constraints on the calculated speech movements, then the user is always likely to be confronted with discontinuities that clearly indicate an error in the position calculation.

[5] For example, if a sensor on the tongue is occasionally susceptible to these pitfalls but also has a large proportion of reliable data then it is possible by means of multiple linear regression to use data of other sensors on the tongue to generate a prediction of the reliable data, and then extend the prediction to all sample points of the unstable sensor. This prediction will probably be within a few millimetres of the correct solution and so can act as a good starting point.

The purpose of the next section will be to show how some typical problems of this kind can be handled. One point that is hopefully readily apprehensible from the above discussion is that any distortion or noise in the measured signals will increase the likelihood of problematic situations. In any realistic measurement situation it is highly improbable that the true solution actually corresponds to all-zero residuals in the raw measured signals, simply because the real-life signals will not be perfectly modelled by the magnetic-field equations. However, because we have a slightly overdetermined system (six transmitter signals to determine 5° of freedom in the sensors) there is a good chance that overall the minimization of the RMS will give a result close to the true solution even if this RMS value is indeed greater than zero. However, the stronger any disturbances are, and the more the RMS value even for the true solution departs from zero, then the greater are the chances that a substantially different solution with similar or even lower RMS exists somewhere in the search space.

AQ: Please check change.

Please revert to original

20.3.2. **Some realistic problems in position calculation**

20.3.2.1. Reducing instabilities by adjusting the raw signal amplitudes

Figure 20.7 illustrates a category of problem that we have commonly encountered in our data. In the recording session from which the example is taken the middle of three sensors on the tongue occasionally appeared to give unstable results. The top left panel in the figure shows movement in the lateral dimension over time for one trial (similar behaviour was observed in the vertical and anterior-posterior dimensions). Just before 1s on the time axis there is a very sharp to-and-fro movement of over 5 mm that is very untypical for speech movements of this sensor.

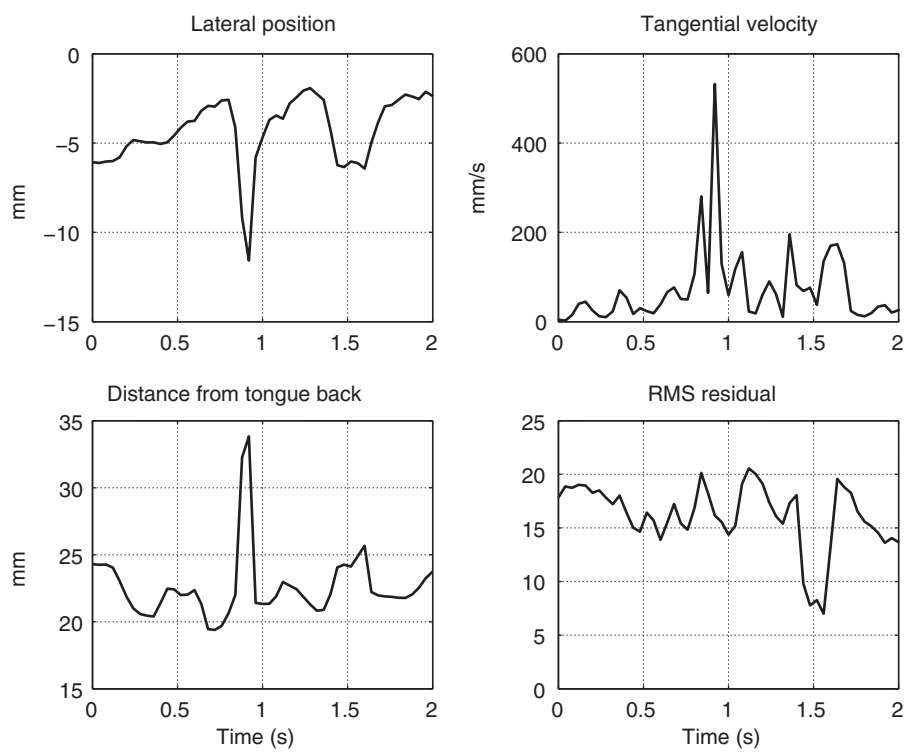

Fig. 20.7 Example of unstable solution of the position calculation.

Correspondingly, the tangential velocity (top right panel) goes up to about 500 mm/s whereas velocities for this sensor would typically not be expected to be much above 200 mm/s. As further evidence of unstable behaviour, the bottom left panel shows the distance of the middle tongue sensor from the back tongue sensor. This shows a large fluctuation of about 10 mm at the same location in time. Overall, it appears very likely that over a period of about 1/10 s the position calculation algorithm has failed to converge on the correct solution. Here this has been presented in terms of eye-balling the data for plausibility, but normally this would be backed up by statistical analysis of the behaviour of the sensor over the course of the recording session.

The bottom right panel shows the RMS residual value; this shows no unusual features in the vicinity of the potential instability, but it is overall quite high (the units of the RMS signal are arbitrary since they are simply related to the units of the AD-convertor that digitized the raw modulated sensor signal, so the statement 'quite high' is just based on what experience has shown to be typical for our system).

The following procedure aims to adjust the raw measured amplitudes so that overall lower RMS values result, in the hope that this will leave less scope for the optimization procedure to veer off towards aberrant solutions. The point of departure for expecting this to have some chance of success is the remark made above that the set of equations is somewhat overdetermined, so in basically stable situations the algorithm will converge on a good solution even if the accompanying minimum RMS residual value is not particularly low, due for example to the sensor picking up spurious signals of some kind or the electromagnetic inductance being slightly different when sensors are attached to the subject compared to the situation when the sensors were calibrated. If these sources of distortion to the signal were purely random noise then nothing more could be done, but if they showed a systematic pattern, i.e., were predictable from the location of the sensor in the measurement field then they could be compensated for, by adjusting the raw amplitudes of each transmitter accordingly. The position calculations can then be repeated based on the adjusted amplitudes. For stable solutions the adjustment can be expected to change the result very little, since one is essentially plugging into the input of the position calculation ~~of~~ the amplitudes generated by the magnetic field equations for the solution found in the first pass. However, for unstable solutions the adjustment can be expected to help tilt the balance in favour of a situation in which the lowest RMS value found during the search procedure actually corresponds to a good solution.

Figure 20.8 shows the residual values for one specific transmitter (Transmitter 3 of the mid-tongue sensor) as a function of all transmitter signals for this sensor. It will be observed that the residual has quite a strong linear relation with respect to several transmitters' signals. Essentially we then use multiple linear regression to predict the residual of each signal from all transmitter signals (it is important when calculating the regression coefficients to eliminate data for which the calculated positions are unreliable, since the residuals of this data would act as a source of noise in the regression analysis).[6]

The result of recalculating the specific example here after adjusting the amplitudes by the predicted component of the residual is shown in Fig. 20.9 where the original values (dashed lines) and the corresponding values for the new solution (solid lines) are overlaid. It will be seen that in

[6] In a preliminary version of this adjustment procedure, we simply predicted the residual of each transmitter from the signal of the corresponding transmitter. This is tantamount to adjusting the gain factor for the corresponding sensor-transmitter pair. But if, as seems likely, disturbances in the signal depend more on position of the sensor in the measurement space than just on the strength of the induced signal, then the overall state of the system can be taken better into account by using all transmitter signals in the prediction.

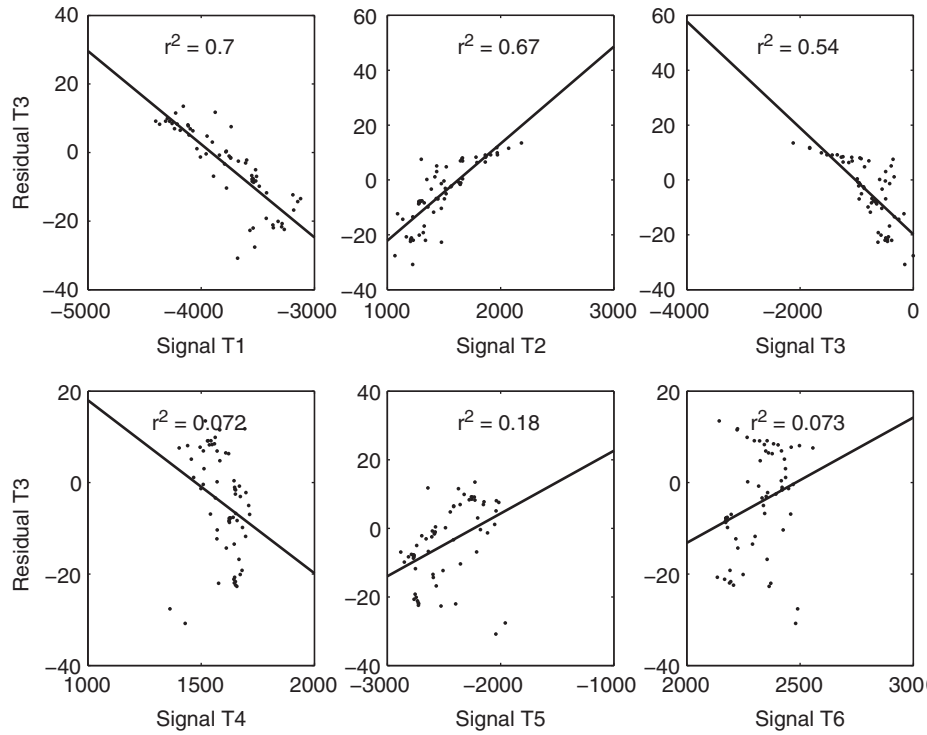

Fig. 20.8 Residual for transmitter 3 plotted as a function of all six transmitter signals.

the panels for position, velocity, and distance from the tongue-back sensor the extreme values in the region of the potential instability have disappeared, but that in other parts of the trial the changes are very small. The substantial decline in the RMS value in the bottom right panel indicates that a large proportion of the original RMS value was systematic and predictable.

20.3.2.2. Reducing instabilities by estimating positions from predicted velocities

The procedures presented in the previous section have proved invaluable in many experiments for reducing the proportion of unstable data. However, some problems may well remain after amplitude correction, Fig. 20.10 showing a case in point. The dashed lines represent the situation at this stage in the processing. A clear discontinuity is still to be observed in the position and velocity trace (again, the lateral dimension has been chosen for illustrative purposes). In this figure the raw data for this sensor (with respect to the six transmitters) has been included as the bottom left panel in order to emphasize that the raw data itself is clearly continuous, i.e., there has not been some low-level measurement error that can explain the present problem. As in the previous section, the 'repair' procedure in the present section has as its first pre-requisite that a reasonable proportion of the calculated positions are basically reliable, and that it is possible to eliminate from the calculations those data points that are likely to be unreliable. The other point of departure is as follows: When the full 5D measurement space is considered then position and orientation of the sensors is a strongly non-linear function of the raw amplitudes induced in the sensors. However, in the short trials typical of many speech experiments (e.g., a few seconds) the sensors will only move through a very small proportion of the total measurement space.

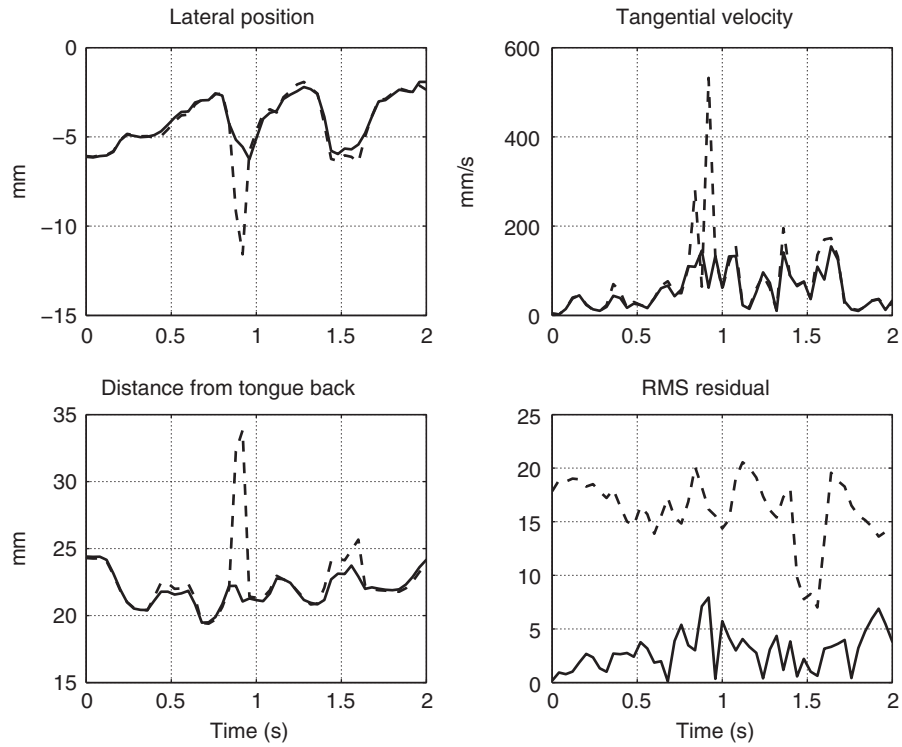

Fig. 20.9 Results for re-calculated positions based on adjustment of raw amplitude data overlaid on the original data (dashed lines: original data; solid lines: re-calculated data).

Therefore, for short stretches of data the relationship between output (position, orientation) and input (induced signals) may be locally linear. This offers the possibility of shortcutting the position calculation based on the non-linear optimization procedure, by using the data where position calculation has been successful to obtain by multiple linear regression a mapping directly from raw amplitude data to the positions. This mapping can then be applied to those data points where the normal position calculation has obviously produced aberrant results. In practice, we obtained better results by predicting sensor velocities from the first derivatives of the raw data. Following integration, the estimated positions in the unstable regions are patched into the data tracks in such a way as to be continuous with the immediately preceding and following stable regions.

The result of this procedure, i.e., the re-estimated data, is given by the solid lines in Fig. 20.10 (overlaid on the original data given by the dashed lines). Clearly, there have been substantial changes in the region of the assumed instability, with the re-calculated version certainly appearing superficially more plausible. The bottom right panel offers a more formal way of assessing whether the re-estimation procedure has produced acceptable results. This overlays the RMS residual of the original position calculation with the corresponding trace for the re-estimated positions. For any set of re-estimated positions (i.e., not just those based on the present procedure) it is absolutely essential to generate the amplitudes predicted for those positions by the field equations and subtract these amplitudes from the measured ones in order to have access to corresponding residual values. These re-estimation procedures may themselves fail, for example, if the identification of reliable and unreliable data was deficient. This may not be immediately obvious if the output is smooth, but if the re-estimated solution is associated with a drastic

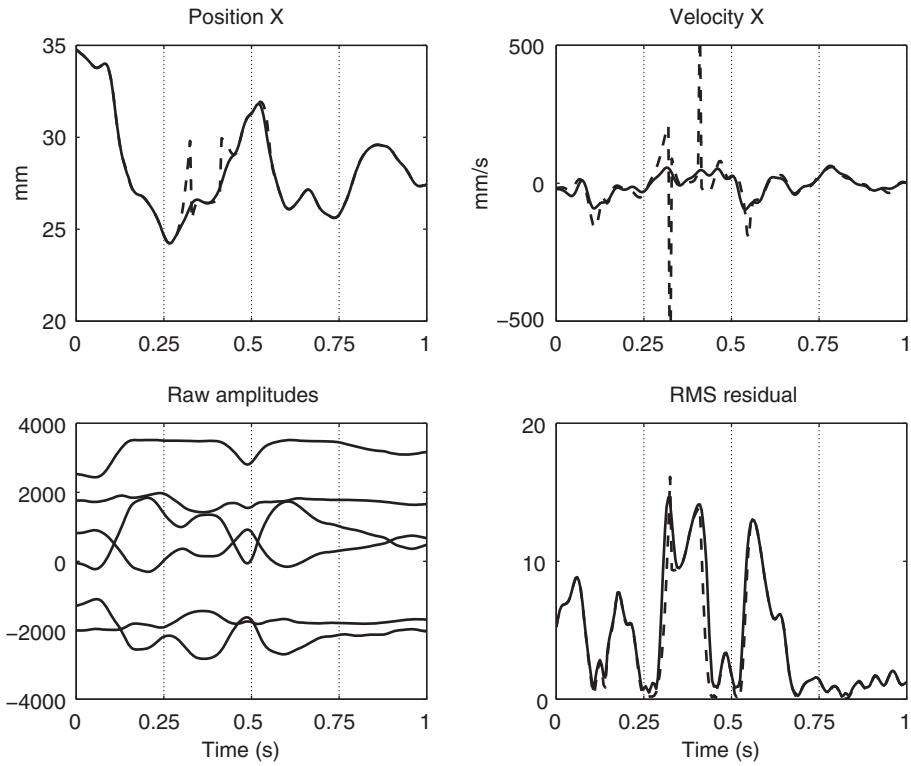

Fig. 20.10 Example of discontinuity in position and velocity data (top panels, dashed lines). Raw data (corresponding to the six transmitters) in the bottom left panel. Data re-estimated by velocity prediction shown in solid lines overlaid on the original data (all panels except bottom left).

increase in the residual values then it must be rejected as unrealistic. In the present case the re-estimation can be regarded as successful because the overlaid traces are almost impossible to distinguish in the figure. This demonstrates with a practical example the point made above with the 'toy' example that solutions with very similar RMS values but widely separated in space can indeed exist. As long as the optimization procedure itself does not incorporate continuity constraints to weight different solutions that may appear equally attractive in terms of RMS then the application of post-hoc procedures of the type outlined here will probably continue to be necessary.

20.4. **Conclusion**

The aim of the present chapter has not been to give an exhaustive review of the state of the art in articulographic measurement procedures, but rather to focus on areas where both the benefit and the bane of the current generation of electromagnetic systems become particularly apparent (for further methodological discussion of current EMA systems see e.g., Zierdt [2007], Yunusova et al. [in press], Kroos [2008], Yunusova et al. [2008]).

AQ: Is it Yunusova et al., 2009 as in the reference list?

Yes

On the plus side, due to the high information density per sensor it is apparent that this technique has the potential to provide data for a wide range of both traditional and innovative experimental paradigms, with the increase in subject comfort particularly useful for large-scale recordings.

On the minus side, the user must be prepared to devote considerable effort ~~in~~ λ assessing the stability of the results of the position-calculation algorithm, and to exploring means of improving the stability of questionable data. Fortunately, the latter usually only constitutes a small proportion of the total amount of data from typical experiments (in our experience perhaps about 10%). But handling this data usually dominates the time required to process a dataset. With regard to the 'repair' procedures outlined above it is worth emphasizing the following points in this concluding section: (1) They depend on a fair proportion of data being basically accurate, and (2) They depend on being able to specify plausible ranges for velocity and inter-sensor distances in speech movements.

In other words, the problems with which we are typically confronted in the current state of the system would be much more intractable if we wanted to measure unconstrained movement of sensors that are not linked among themselves (at least loosely) by being attached to biological tissue. For example, the tongue-tip and tongue-dorsum are sufficiently independent of one another for it to be interesting to study their coordination. But the range of positions they can adopt relative to one another is much more constrained than, for example, the index fingers of the left and right hand.

Nonetheless, it would, of course, ultimately be desirable and more elegant to avoid 'repair' procedures altogether (i.e., procedures that 'patch' after the event), by building more sophisticated constraints a priori into the optimization problem itself. This will be an interesting task for future developments.

References

Anderson A, Bader M, Bard E, Boyle E, Doherty GM, Garrod S, Isard S, Kowtko J, McAllister J, Miller J, Sotillo C, Thompson HS and Weinert R (1991). The HCRC Map TaskCorpus. *Language and Speech*, **34**, 351–366.

Burnham D, Lau S, Tam H and Schoknecht C (2001). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. *Proceedings of AVSP 2001*, pp. 155–160.

Craig MS, van Lieshout PHHM and Wong W (2008). A linear model of acoustic-to-facial mapping: model parameters, data set size and generalization across speakers. *Journal of the Acoustical Society of America*, **124**(5), 3183–3190.

Hoole P and Nguyen N (1999). Electromagnetic articulography in coarticulation research. In WH Hardcastle and N Hewlett, eds. *Coarticulation*: *Theory*, *Data and Techniques*, pp. 260–269. Cambridge University Press.

Hoole P and Hu F (2004). Tone-vowel interaction in standard Chinese. *Proceedings of the International Symposium on Tonal Aspects of Languages*: *Emphasis on Tone Languages*, *Beijing*, pp. 89–92.

Kaburagi T, Wakamiya K and Honda M (2005). Three-dimensional electromagnetic articulography: a measurement principle. *Journal of the Acoustical Society of America*, **118** (1), 428–443.

Keating P, Baroni M, Mattys S, Scarborough R, Alwan A, Auer E and Bernstein L (2003). Optical phonetics and visual perception of lexical and phrasal stress in English. *Proceedings of the 15th International Congress of Phonetic Sciences*, *Barcelona*, pp. 2071–2074.

Kröger B, Pouplier M and Tiede M (2008). An evaluation of the aurora system as a flesh-point tracking tool for speech production research. *Journal of Speech Language and Hearing Research*, **51**(4), 914–921.

Kroos C (2008). Measurement accuracy in 3D electromagnetic articulography (Carstens AG500). In R Sock, S Fuchs and Y Laprie, eds. *Proceedings of the Eighth International Seminar on Speech Production*, *Strasbourg*, *2008*, pp. 61–64.

Ling ZH, Richmond K, Yamagishi J and Wang RH (2008). Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. *Proceedings of the Interspeech Brisbane*, *2008*, pp. 573–576.

Mixdorff H and Charnvivit P (2004). Visual cues in Thai tone recognition. *Proceedings of theInternational Symposium on Tonal Aspects of Languages*: *Emphasis on Tone Languages*, *Beijing*, pp. 143–146.

Munhall KG, Jones JA, Callan DE, Kuratate T and Vatikiotis-Bateson E (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science*: *A Journal of the American Psychological Society /APS*, **15**(2), 133–137.

Yehia H, Kuratate T and Vatikiotis-Bateson E (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, **30**(3), 555–568.

Yunusova Y, Stanley J and Green JR (2008). Distinguishing place of consonant articulation using the aurora system. ASA meeting Paris Summer 2008. Paper 4pSCb35. *Journal of the Acoustical Society of America*, **123** (5, Pt. 2), p. 3739.

Yunusova Y, Green J and Mefferd A (2009). Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech Language and Hearing Research*, **52**, 547–555.

Zeroual C, Hoole P, Fuchs S and Esling JH. (2007). EMA study of the coronal emphatic and non-emphatic plosive consonants of Moroccan Arabic. *Proceedings of the 16th International Congress of Phonetic Sciences*, *Saarbrücken*, pp. 397–400.

Zierdt, A. (2007). EMA and the crux of calibration. *Proceedings of the 16th International Congress of Phonetic Sciences*, *Saarbrücken*, pp. 593–596.