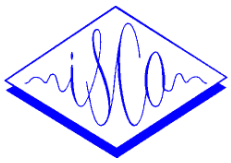

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

EXTRACTING ARTICULATOR MOVEMENT PARAMETERS FROM A VIDEODISC-BASED CINERADIOGRAPHIC DATABASE

Mark K. Tiede and Eric Vatikiotis-Bateson
ATR Human Information Processing Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-03, Japan

ABSTRACT

The recent transfer of a selection of cineradiographic movies of the vocal tract to the laserdisc format has now made the application of standard video processing techniques to such data feasible on inexpensive microcomputers. This research explores some of the possibilities inherent in this new look at old data: we have developed software capable of automatically reducing noise, detecting air-tissue boundaries, and tracking articulator displacements (e.g. jaw, lips, tongue) through successive frames. Our goal is to characterize tongue dynamics in terms of changes over time to a normalized parameterization of the tract surface profile. This information will complement other modes for investigating speech production for which either dynamics or surface profiles are unavailable (MRI, magnetometer).

MOTIVATION

The full sagittal view of vocal tract articulators during running speech provided by cineradiography remains unsurpassed by more modern techniques, and it has been a staple of phonetics research for many years since the pioneering research of Chiba & Kajiyama (1941). For example, the influential work of Fant (1960) on modeling vocal tract acoustics relies extensively on analysis of cineradiographic data, as do the articulatory models of Mermelstein (1973) and Maeda (1979). Literally hundreds of cineradiographic studies of speech production have been performed: a database compiled by Dart (1987) lists 282 different sources (though not all are movies).

Making quantifiable use of such data however has until now depended on laborious and error-prone hand tracing from individual movie frames, which has imposed practical limitations on both the scope and the kinds of measurements possible. The volume of data amassed at a typical 50fps capture rate precludes hand analysis of every frame except for very short sequences, so measurements have normally been taken from key frames (corresponding to articulatory targets) chosen by visual inspection.

This is unsatisfactory for two reasons. Apart from the subjective character of hand tracings and the criteria used for key frame selection, measurement errors may arise from variability introduced by scintillation of the X-ray source, mechanical jitter in the recording camera, or image noise resulting from film processing. Second, and more importantly, the unique aspect of cineradiographic data is the visualization it provides of time-varying midsagittal tract surface shapes resulting from synergistic interaction between articulators, but measurements sampled at target-determined intervals lose or at best distort the dynamics of these changing shapes. Such target or key frame driven analysis is analogous to making a railway journey in which one looks out the window only at station stops, missing all the details of the intermediate journey, which are arguably the most interesting.

Both these issues can be addressed by automating the analysis process. Computer control permits measurement with reproducible consistency, manipulation using standard image processing techniques to reduce noise and enhance visual features of interest, and application of some analysis algorithm to all available data frames. The recent transfer of a representative selection of high quality radiographic movies to video laserdisc (Munhall, Vatikiotis-Bateson, and Tohkura 1994) has now made such automated analysis possible on widely available microcomputers.[1] The videodisc format permits software to individually address and precisely control images corresponding to original movie frames. Because disc players are capable of stable 'freeze-frame' video output, an analysis program can proceed at its own pace through the frame digitization, image enhancement, and measurement cycle before stepping the player to the next frame.

There are three desirable consequences of the transfer of X-ray data to this format. The first is that digital image processing techniques can be applied as appropriate, resulting in enhanced images of potentially better quality than the original movie. Secondly, because measurements can be obtained automatically at a far greater rate than previously possible, all available frames can be processed, instead of the measurement of key frames feasible by hand analysis. Finally, because movies in this format are much easier to work with, experimentation is easy, and we report our efforts in the hope that others will attempt similar exploration.

HARDWARE

The functional components of our analysis setup are shown in Figure 1, and the specific equipment that we use is listed in Table 1.

Computer:	Apple Macintosh Quadra 950
Digitizer:	RasterOps 24STV
Laserdisc player:	Pioneer LD-V800

Table 1: Hardware used

We use a standard microcomputer arrangement augmented by a digitizer board, a second monitor, and a videodisc player supporting software commands issued over a serial (RS-232) link. A large hard drive supplemented by an auxiliary magneto-optical cartridge drive is used for storing digitized images.

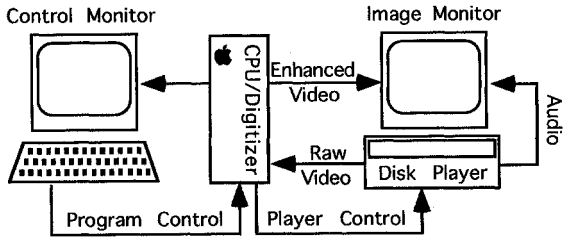

Figure 1: Hardware setup

The digitizer supports 'play through' video at the standard 30fps rate, which is useful for disk cueing operations. Digitization occurs at a depth of 24bits per pixel, which is converted in software to 8bits per pixel grayscale. Using this equipment it takes approximately one half second to digitize a full video frame.

All player functions are programmable, and are transmitted via a handshaking protocol over a serial link between the player and the computer modem port. The player supports the constant angular velocity format (CAV), in which one video frame is encoded in each circular track on a videodisc. Use of this format permits stable video output at any playback speed, including while paused ('freeze-frame'), making possible the asynchronous stepping required for automated analysis.

SOFTWARE

Our software development was driven by the following design criteria: we wanted program control of both videodisc player and digitizer functions without being limited to a specific hardware configuration, flexibility permitting easy modification and implementation of image enhancement and measurement algorithms, and the means to export both acquired images and measurement results to other application programs. Although the software continues to evolve in the direction of improved modularity, we have succeeded in building a tool that supports our primary goal of experimenting with different approaches to image analysis.

Hardware independence was achieved in the case of the digitizer by driving it through the Apple QuickTime operating system extension; because the program is thus isolated from digitizer-specific details, it can be used with any QuickTime-compatible board. Video player control was implemented by organizing model-specific control calls into separately compiled modules communicating through a standardized interface to the program. In this way the program can be used with different types of players by selecting an appropriate interface module.[2]

Basic functionality provided by the program includes player cueing support, selection and scaling of a subsection of the video signal to use for acquisition, control over various aspects of digitization, and facilities for saving and recovering particular configurations. Digitized images may be saved in either the PICT or TIFF formats; measurement data is saved to spreadsheet-compatible files.

Image enhancement routines are separately compiled modules loaded as needed by the program, and applied to acquired images as directed by the user. This approach permits adding new enhancement capabilities without requiring changes to the main program. Implemented modules include algorithms for image histogram normalization, jitter correction, edge detection, and various kinds of filters.

Measurements can be obtained either interactively or automatically. In interactive mode the user hand positions measurement objects (points, lines, Bézier curves) to align them with features of interest in selected images. Automatic measurement is based on detecting the position of a significant change in pixel intensity along a line superimposed on successive image frames. The user positions as many of these lines as desired on an initial image, in any orientation, and the location of the intensity boundary is determined and recorded for all subsequent images in the range to be analyzed. Figure 2 shows an initial image with a set of superimposed measurement lines. The software is written in C, and is available for testing purposes to any not-for-profit institution upon written request.

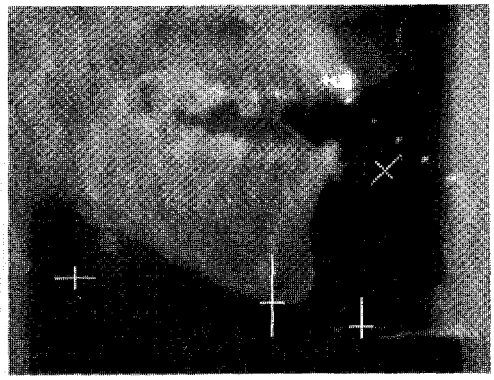

Figure 2: Sample image with superimposed measurement lines

DATA

There are 25 X-ray films currently available to us on videodisc from 14 different subjects, totaling about 55 minutes of usable footage. Each original movie frame was digitized as a corresponding video frame. Because the film speeds exceed the 30fps standard video frame rate the concurrently recorded audio was stretched to using a pitch-preserving technique to preserve synchronization (see Munhall et al. 1994 for a description of the transcription process).

Twenty-four of these films were made at the Département de Radiologie de l'Hotel-Dieu de Québec, Québec, Canada under the direction of Dr. Claude Rochette (Rochette 1973). The subjects were native speakers of either Canadian French or English, speaking single repetitions of short sentences. Data were recorded at a rate of 50 frames per second. The 25th film was made in 1962 by Dr. Sven Öhman and Prof. Kenneth Stevens at the cineradiographic facility of the Wenner-Gren Research Laboratory at Nortull's Hospital, Stockholm, Sweden. The subject was Kenneth Stevens; the stimuli were two sentences and a series of nonsense mono- and bi-syllabic utterances exercising the subject's vowel inventory. The recording frame rate was 45 fps. Documentation of the recording techniques used for this film and a thorough analysis of the data are available in Perkell (1969).

All of the films show the subject in profile: the vocal tract is visible over a vertical range from the level of the larynx to the maxilla, and from the rear pharyngeal wall to the lips. The Öhman and Stevens film shows a slightly larger range, in which the first six vertebrae and some sublaryngeal structure are visible; in addition this film shows a line of reference pellets spaced one centimeter apart that can be used for calibrating measurements.

ANALYSIS ISSUES

While the films are of high quality and have been transferred to video format with minimal loss of resolution, data of this type present certain inherent problems for computer-based analysis. In particular, scintillation of the original X-ray source is superimposed as randomly distributed high-frequency noise. Low-pass filtering reduces its effects, but at the cost of degrading the high-frequency edges associated with articulators in motion. To some degree median filtering and thresholded neighborhood averaging techniques can be usefully applied (see e.g. Gonzalez & Wintz 1987), but the intensity boundary detection algorithm remains sensitive to residual noise of this type.

Another kind of noise affecting the recurrent measurement of a feature of interest through a series of images is spatial dislocation of the frame of reference (jitter) caused by mechanical vibration in the recording camera or other factors. Jitter observable in the data is not sufficiently periodic to permit effective post-measurement filtering. However, by use of a stable reference edge consistently present through the series the dislocation can be tracked along that dimension, and its effect automatically subtracted from the measurements obtained.

A different kind of problem is the difficulty of tracking articulator positions while obscured by structures opaque to X-rays. The jaw, the hard palate, and dental fillings in particular can obscure portions of the tongue surface in a given frame. But when analyzed as part of a series the location of an obscured contour can be estimated subject to constraints extrapolated from previously observed locations and some knowledge of the physical limitations of the system (e.g. location of the palate roof represents an upper limit for possible tongue position). Kass, Witkin, and Terzopoulos (1988) have implemented an approach of this type using energy-minimizing splines ("snakes") to successfully track lip aperture through successive frontally oriented video images of a speaker's mouth, and we are currently attempting to adapt the technique to tongue surface tracking.

Even when the tongue is fully visible it can sometimes be difficult to define its midsagittal position: grooving especially in the posterior dorsal position may cause more than one surface to be shadowed by the X-ray source. Recognizing this as a potential problem, Öhman and Stevens outlined the midline of the subject's tongue and lips with an X-ray impeding barium compound, which is generally sufficient in that film for distinguishing the midsagittal shape. In all the films, however, the midline of the tongue can be detected by augmenting the edge or threshold algorithm used to reject all but the most central of the detected contours.

APPLICATION

The same wealth of detail that makes the cineradiographic images fascinating to watch can be overwhelming for any attempt to quantify salient articulator movement. One straightforward approach is to track the spatial displacement of consistent visual features associated with an articulator, deriving a pattern of movement analogous to point source trajectories obtainable using X-ray microbeam or magnetometer techniques. As a comparison benchmark for testing such derived trajectories we were able through the cooperation of Professor Stevens to record him speaking the same stimuli in a 1994 magnetometer experiment that elicited articulator trajectories paralleling his original film. In this experiment transducers were mounted on the upper and lower lips, the jaw, and at four locations along the midline of the tongue (two additional points were tracked for head correction purposes). Horizontal and vertical components of transducer movement data were sampled at 200Hz. Audio was sampled simultaneously at 10KHz. Figure 3 shows a comparison of the magnetometer-derived vertical component of jaw trajectory and the results of automatically tracking the lower mandible along the reference line shown in Figure 2: after median filtering, histogram stretching, and jitter correction, the measurement for each frame was determined from the location of the pixel intensity boundary associated with the jaw.

In the figure the trajectories have been temporally aligned using the initial and final jaw lowering gestures. Although separated by more than thirty years and derived quite differently, the comparison shows considerable similarity in gestural timing through the sequence (regression of the X-ray values against the magnetometer trajectory gives R2 = .45). Apart from demonstrating the validity of the automated analysis technique, this suggests that such similarity can be exploited as a general means of temporally aligning X-ray images to related trajectory data from other sources.

Why would one wish to do so? Speech production data

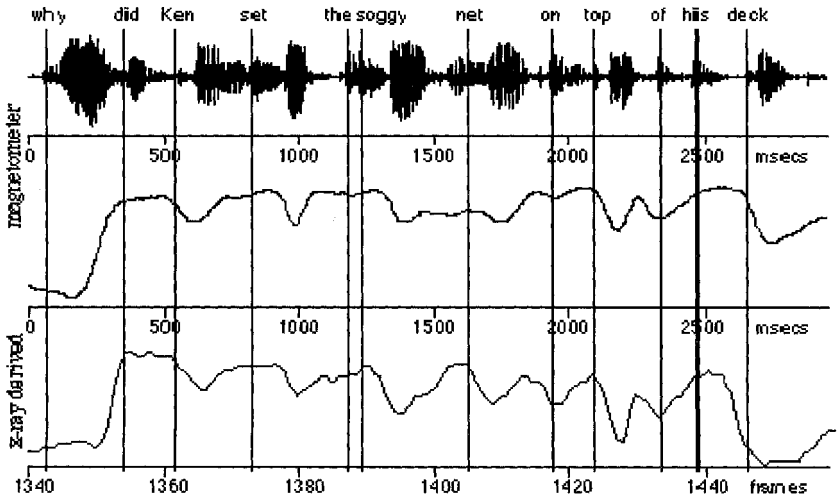

Figure 3: Comparison of magnetometer and X-ray derived jaw trajectory (vertical component).

collection techniques are currently of two types: one permits real time tracking, but limited to a few coplanar points (microbeam, magnetometer); the other gives full volume vocal tract images, but limited to static, sustainable configurations (MRI, CT). Cineradiography occupies a middle position, but since concerns over radiation hazards preclude its continued use, studying the data already collected offers the potential for interpolating between the deficiencies of these other techniques. In particular, the dynamically varying tongue contour extracted from an image sequence could be used to extrapolate from fitted tongue points to tongue root position, not normally accessible to point source techniques. Alternatively deformation of an MRI-derived representation of tract volume between known target configurations could be driven by the tongue shapes extracted from X-ray images.

Effective use of such contours, however, extracted in terms of detected air-tissue boundaries, presupposes that they have been converted to an appropriate parameterized description of the tongue shape. Characteristics of such a parameterization should include the ability to cope with the complex tongue shapes that arise from braced tongue postures. They should also be derivable on a frame-by-frame basis. Moreover, the parameterization itself should facilitate dynamical description.

Previous efforts to parameterize tongue shape include the PARAFAC analysis of Harshman, Ladefoged, and Goldstein (1977), and Jackson (1988), and the quadratic fitting of Hashimoto and Sasaki (1982). Both Harshman et al. and Jackson obtained results showing that just two factors were adequate for describing English vowel shapes (though Jackson's work suggests that an additional factor is necessary to account for inter-language variability). Although this represents tract configuration in an especially concise form, the factors conflating tongue shape and position are defined with respect to all sampled configurations. Therefore, they cannot be generated as the representation of a particular curve instance. For our purposes a potentially more useful approach is that followed by Hashimoto and Sasaki, who fit a quadratic describing tongue shape to points sampled along a superimposed grid. However, while adequate for the vowel target configurations they chose to analyze, second-order curves are clearly inadequate to describe the more complex shapes assumed by the tongue in consonantal gestures involving tongue tip bracing against the alveolar ridge (see Stone 1990 for data and discussion). Figure 4 below illustrates such a configuration.

In our own work we are investigating the use of Bézier curves (see e.g. Farin 1990) as a suitable parameterization type. Under this approach a curve is specified by the positions of control points defining an enclosing polygon, which has the useful property that a change in the position of any control point is reflected globally along the entire approximated curve. If some offset along the curve is thus treated as an analog to a point on the original tongue, then as the control points shift to match the approximation to the observed shape, that reevaluated offset tracks the new position of the tongue point. This permits comparison to tongue pellet trajectory data from microbeam and magnetometer sources.

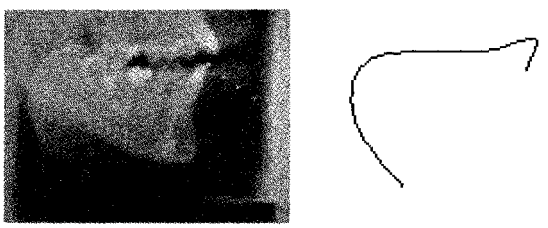

Figure 4: /l/ ("blood"); image and fitted curve (6th order)

SUMMARY

The transfer of vocal tract cineradiographic data to the videodisc format has made possible new kinds of analysis of such data. It is hoped that efforts sketched here will encourage others in similar experimentation, and provide impetus for transferring additional X-ray movie archives to the videodisc format.

ACKNOWLEDGMENTS

The authors thank Ken Stevens for serving as the subject once again, 32 years after his original stint, Vincent Gracco and Elliot Saltzman for experimental assistance, and Kevin Munhall for suggestions and encouragement.

REFERENCES

Chiba, T. and M. Kajiyama (1941) The vowel: its nature and structure, Tokyo-Kaiseikan, Tokyo.

Dart, S. N. (1987) "A bibliography of X-ray studies of speech," UCLA Working Papers in Phonetics, 66, 1-97.

Fant, G. (1960) Acoustic theory of speech production, Mouton, 's-Gravenhage.

Farin, G. (1990) Curves and surfaces for computer aided geometric design, 2nd Edition, Academic Press, San Diego.

Gonzalez, R. C. and P. Wintz (1987) Digital Image Processing, 2nd Edition, Addison-Wesley, Reading, MA.

Harshman, R. A., P. N. Ladefoged and L. Goldstein (1977) "Factor analysis of tongue shapes," Journal of the Acoustical Society of America, 62, 693-707.

Hashimoto, K. and K. Sasaki (1982) "On the relationship between the shape and position of the tongue for vowels," Journal of Phonetics, 10, 291-299.

Jackson, M. T. T. (1988) "Analysis of tongue positions: Language-specific and cross-linguistic models," Journal of the Acoustical Society of America, 84, 124-143.

Kass, M., A. Witkin and D. Terzopoulos (1988) "Snakes: active contour models," International Journal of Computer Vision, 1, 321-331.

Maeda, S. (1979) "Une modele articulatoire de la tongue avec des composantes lineaires," JEP, GALF, 10emes, 152-164.

Mermelstein, P. (1973) "Articulatory model for the study of speech production," Journal of the Acoustical Society of America, 53, 1070-1082.

Munhall, K. G., E. Vatikiotis-Bateson and Y. Tohkura (1994) "X-ray film database for speech research," Journal of the Acoustical Society of America, 95, No. 5, Pt. 2 (Supplement), 2822.

Perkell, J. S. (1969) Physiology of speech production: results and implications of a quantitative cineradiographic study, MIT Press, Cambridge, MA.

Rochette, C. (1973) Les groupes de consonnes en Français, Les Presses de l'Université Laval, Québec.

Stone, M. (1990) "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," Journal of the Acoustical Society of America, 87, 2207-2217.

NOTES

1 The X-ray film database on laserdisc is freely available to research institutions; for details see Munhall et al. 1994.

2 To date, control modules exist for the Pioneer laserdisc player models LD-V800 and CLD-V2400, and the Sony Hi-8 videotape player model EVO-9650.