1040

International Symposium on Simulation, Visualization and Auralization for Acoustic Research and Education

ASVA 97

2-4 April 1997 Tokyo, Japan

VISUALIZATION OF SPEECH PRODUCTION: FOR BETTER AND FOR WORSE

Eric Vatikiotis-Bateson **Mark K. Tiede**
ATR Human Information Processing Res. Labs., Japan
Philip E. Rubin[1] **Christian Benoît**
Haskins Laboratories, USA *Institut Communication Parlée, France*

Abstract. The role of visualization in speech production research is discussed from the perspective of efforts at our respective laboratories to develop research tools appropriate to the analysis and representation of behavior that is inherently dynamic, multidimensional, and only partially visible to external view. Accelerating advances in technology over the past few years have brought us much closer to achieving this goal, while providing the means to pose quite different questions than before. Despite the obvious benefits, the scale and speed of the recent advances in visualization, such as animation, also pose problems at a number of levels. While such "growing pains" can be managed, they cannot be ignored.

During the past twenty years, improvements in computer hardware and software have radically changed our approach to the study of speech production. Although this is a common story for many fields, the case of speech production is interesting because its neuromotor and musculoskeletal underpinnings, which can be observed in other biological systems, are complex and very difficult to observe directly. Inevitably, then, our conceptual and analytic approaches to speech production have been shaped by our ability to visualize the phenomena.

We provide an historical overview of our efforts to produce visualization tools which are simultaneously powerful enough to achieve the qualitative, quantitative, and conceptual goals associated, for example, with the estimation of time-varying vocal tract parameters and subsequent acoustic synthesis from physiological data and linguistic control strings. This and other examples are drawn from the long-term effort at Haskins Laboratories and more recently at ATR (Japan) and the ICP (France) to develop accessible, yet powerful, analytic and representational tools for articulatory and audiovisual modeling.

We conclude with a brief discussion of two recent developments that we believe are already responsible for improved communication among engineers, psychologists, physiologists, and linguists and subsequently for an explosion of elaborate models and ingenious preliminary analyses of complex data structures. Unfortunately, there is always a lag between technological and conceptual advances. Today, visualization is so seductive — e.g., the animated, color 3D demonstration — that it threatens to become an end unto itself, rather than the powerful means towards the conceptual advancement which is sure to follow.

A PATH TO THE PRESENT

Haskins. For the purpose of this discussion, two approaches to speech production research conducted at Haskins are pertinent. One concerns the substantial empirical and methodological challenges related to recording, processing, and analyzing multiple physiological signals associated with the moving vocal tract. The explicit goals of this approach have been to identify and characterize linguistically-relevant articulatory events at one or more levels of behavioral observation — e.g., the articulator kinematics and the associated muscle activity (EMG) — and to use them to distinguish between variable and invariant aspects of the resulting speech behavior. The other approach has been

less-concerned with the details of observable production data and more concerned with developing a coherent scheme for generating a plausible output from a manageable set of vocal tract or even more cognitive control parameters. Historically both approaches owe much to the development of the Haskins Pattern Playback [4]. The Pattern Playback could synthesize perceptually contrastive stop consonants (e.g., /b,d,g/) in stop+vowel sequences simply by altering the frequency sweep of formant patterns. In addition to demonstrating the perceptual relevance of formant transitions, their dynamism pointed up the importance of studying the moving vocal tract.

Normal speech production entails the continuous interplay of all regions of the vocal tract and consequently all contributing articulators. Ideally then, we would like to record and measure the time varying behavior of the entire vocal tract. But there are serious problems preventing this. First, with the exception of the lips and, indirectly, the jaw, the vocal tract is hidden from external view. Second, changes in vocal tract shape are brought about by the coordinated, but distributed, activity of usually deformable articulators. Finally, articulator motions are small compared to other biological motions and can be quite fast (e.g., tongue tip trills, bilabial bursts), thus requiring good spatiotemporal resolution of the movement transduction system.

The type of system that comes closest to meeting all three requirements is full-head X-ray ciné. Although available since the 1920's, the health risks are well-known and it is only in the last few years that image processing and analysis techniques have developed that make it possible to do feasible frame-by-frame analysis of ciné film [20]. Data collection techniques at Haskins and elsewhere followed the more manageable and safer course of flesh-point measures transduced optoelectronically, piezoelectrically, or electromagnetically [16].

The biggest methodological problem facing the flesh-point approach was how to best approximate the behavior of the vocal tract structures using a limited number of channels. Even assuming that only two-dimensional motions within the midsagittal plane are necessary, two hardware channels must be dedicated to each flesh-point. Then, how many channel-pairs are needed to capture the motion of deformable structures such as the tongue, and even more troublesome, what are their ideal locations? Add to this set more channels for recording the speech acoustics and relevant muscle activity (EMG) and suddenly 30-40 channels are needed at large aggregate throughputs.

In the 1970's and 1980's the Haskins data acquisition system was limited to analog recordings of 10 channels or less. In order to record both articulator kinematics and muscle EMG, often only one dimension of motion could be recorded [8] and, even in purely kinematic studies, it was standard practice to analyze only one dimension of motion [21]. As a result, multi-channel data were displayed synchronously as one dimensional time series. Such a display made with a Haskins multichannel display program (HADES) is shown in Figure 1.

Such displays encouraged the practice of measuring temporal offsets (and later relative phasing) between events identified visually from the display screen. Thus, for many years, peaks and valleys of position time series corresponding to maximal excursion of the articulator flesh-point were the criterial events for subsequent analysis (see Figure 1). While instructive, this practice had a number of drawbacks. In addition to the already mentioned identification of criterial events from a screen display, only a very small number (1-3) of data points were used for any given channel of a particular experimental token. All other points were ignored. Both of these problems were exacerbated when examining muscle EMG data, whose bandwidth and noise interference both tended to obscure the shape of the signal when represented as a time series. Yet, criterial events — typically, onset and peak of activity, rarely duration — were defined on the basis of screen displays. Analytically, the most common treatment of the small numbers of measures per trial was to do

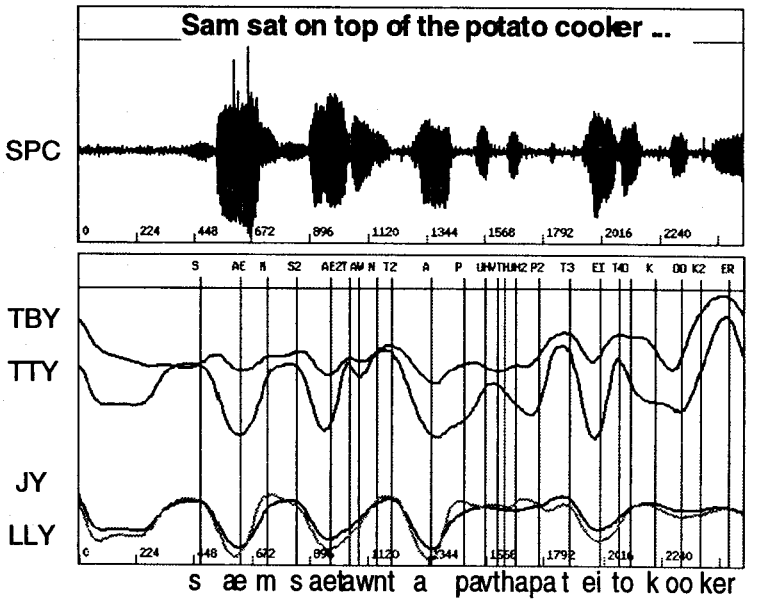

Fig. 1. Acoustic and vertical position time series for tongue blade (TBY), tongue tip (TTY), jaw (JY), and lower lip (LLY).

variance analysis (e.g., ANOVA) of paradigmatic contrasts in phonetic structure, stress, and or speaking rate. Null hypothesis testing on miniscule data samples was used to address lofty problems linking the kinematic and physiological levels of observation — e.g., the equipotentiality of motor control, the span and scope of motor planning, the existence of coordinative structures, and numerous issues falling under the rubric of coarticulation. With so few data points from each type of signal, is it any surprise that clear patterns linking the two levels of observation were hard to find? During this period, use of such visualization and analysis techniques almost insured a distorted sense of the vocal tract behavior.

During the mid 1980's, interest shifted to the analysis of how relations among variables were patterned through time. This coincided with the availability of relatively inexpensive computers with significantly faster processors and larger memories. In particular, the interest in representing articulator behavior in the phase plane led to use of lissajous and other two-dimensional displays (see Figure 2). Although static, such displays made it possible to visualize articulator motion paths at least in the midsagittal plane and to make inferences about the dynamics of the moving system. Improved visualization quickly raised concerns about system geometry, such as the proper orientation of the coordinate reference frame [27] and whether or not the midsagittal plane was sufficient for characterizing vocal tract motion [22]. Being able to visualize aspects of the underlying dynamics qualitatively gave impetus to the search for dynamic control parameters and eventually their quantification.

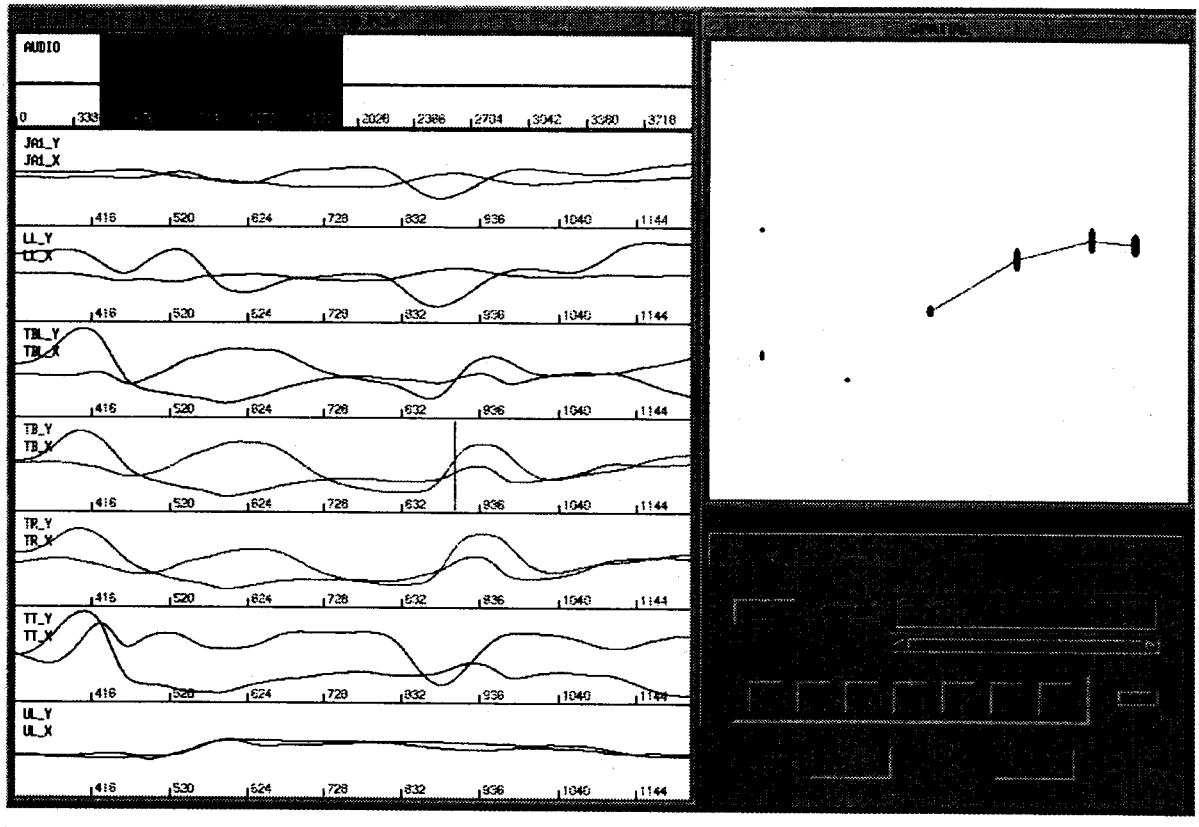

Fig. 2. User screen from MAVIS for displaying animated time-series and 2D plots. Shown here are motion paths for markers on the jaw (1), lips (2), and tongue (4).

In the late 1960's, Paul Mermelstein and colleagues at Bell Laboratories attacked the problem of mapping between vocal tract configurations and appropriate area functions for shaping the acoustic output [12]. In the tradition of the Pattern Playback, the Haskins Articulatory Synthesizer (ASY) was implemented [15] to provide a research tool for rapid synthesis and perceptual evaluation of speech acoustics through simple graphical manipulation of idealized vocal tract control parameters such as the position of the tongue ball and tip, the jaw, the lips. and the hyoid bone. Using key frame animation, a sequence of target configurations could then be interpolated and the acoustics synthesized by addition of a controllable periodic source. Of course, in order to achieve rapid synthesis using the computer technology of the time, many compromises and simplifications had to be made [15]. Figure 3 shows the vocal tract outline and control parameters for ASY.

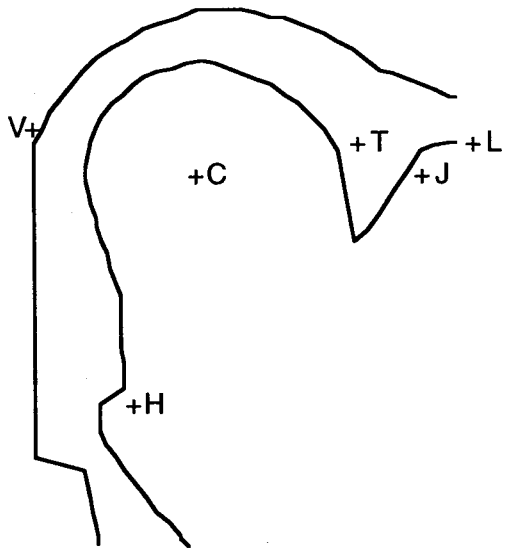

Fig 3. Vocal tract outline from ASY showing articulator control parameters for hyoid (H), velum (V), tongue center (C) and tip (T), jaw (J) and lips (L).

Although the realism of vocal tract configurations depended largely on the experience of the user with real data — e.g., by watching full-head X-ray ciné, analysis of articulatory data — graphics programs were developed (ACE: the Articulatory Control Editor) that allowed synthetic and observed articulator motions to be qualitatively and quantitatively compared. This effort acknowledged the need to provide tools for serious examination of the articulatory kinematics, and it coincided with the intensifying effort to fit second-order equations of motion to articulator behavior [9, 10, 19]. Such equations were intriguing because their kinematic variables (position, velocity, and acceleration) could be derived directly from observation of the data and their dynamic parameters (mass, vis-cous damping, stiffness) could be ascribed inferentially to the underlying neuromotor and biomechanical structures.

Interestingly, this development enhanced interest in inverse estimation of hypothetical control parameter values, without reviving the previous interest in mapping the relations between muscle activity and subsequent motion. The major model of articulator control in speech was the Task Dynamic Model [17]. It incorporated assumptions about physical and functional constraints on articulator behavior as suggested by perturbation studies [8], and focused one level of parametrization on the easily visualized notions of the location and degree of constriction assigned to various vocal tract parameters, such as tongue body, tongue tip, oral aperture, etc. — i.e., the same parameters that could be visually controlled in the articulatory synthesizer.

Task dynamics lacked a front-end controller based on the physical properties of the system (still largely unknown). Front-end control of the task dynamics was provided by the linguistic gestural model of Browman and Goldstein [3]. Although the biomechanics and neuromotor system were by-passed, this model provided the first computational bridge between abstract linguistic entities and testable estimations of articulatory behavior (see Figure 4).

ATR. Whereas the modeling efforts at Haskins were heavily tempered by visual representations of vocal tract structure and behavior, and somewhat less so by quantitative techniques [cf. 7, 13], the emphasis at ATR initially was almost entirely computational and depended little on qualitatively derived inferences about the behavior. Much effort was expended visualizing the block diagram of the control structure, not its activity. Since the contents of the structure were typically expressible in mathematical form, visual verification of the inner workings of the model were not necessary. The only thing necessary was a good simulation of the output. Our colleagues at ATR showed almost no interest in developing interactive user interfaces, which makes good sense given the time scale of the analyses being conducted in the early 1990's. For example, construction of the physiology-based model of speech production involved nonlinear correlation of all data points, not just a few chosen from a graphics display as done previously at Haskins. Estimation of the forward dy-

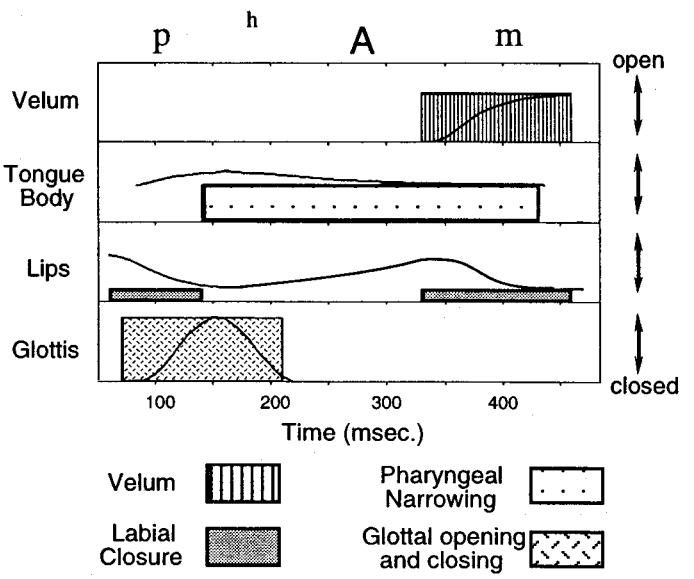

Fig. 4. Gestural score controlling the magnitude and duraton of pharyngeal constriction and velic, labial, and glottal aperture parameters for the word "palm".

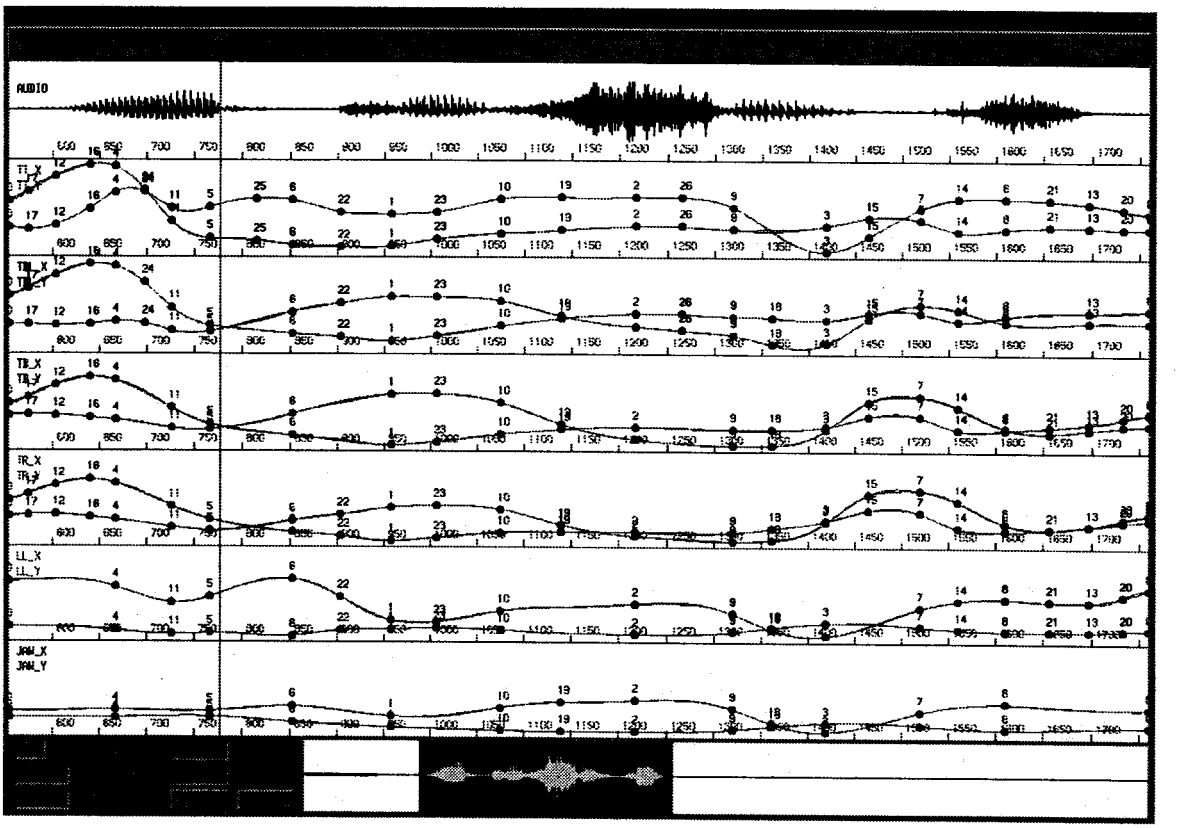

Fig. 5. Display screen for via point estimation tool (VELMA) applied to 2D tongue, jaw, and lip kinematics.

namics linking muscle activity and subsequent articulator motion was done using artificial neural networks in which every possible connection was computed between the 40-50 muscle activity and kinematic channels of the input layer, the units of a hidden layer, and the 15-20 channels of the output layer [6]. This was a major improvement over our previous quantitative methods because the dynamics were being estimated from the mapping between observed events at physiological and kinematic levels, not simply being inferred from the kinematics using our assumptions about the control variables. However, enormous computing power and days rather than minutes were required to generate answers. Also, the complexity of these massively parallel analyses typically required program recompilation every time the data set was altered in any way; this further discouraged the development of user interfaces.

Nevertheless, in collaborating on problems such as via point estimation [25] in which a research paradigm devised for handwriting recognition was applied to speech production [26], it was difficult to separate inadequacies of the model from difficulties posed by adapting it to a different type of biological behavior. In such a case, it was essential to develop interactive tools for extracting via points and displaying the results. Figure 5 shows a typical multichannel display of articulator data and the movement paths reconstructed from the via point analysis using VELMA. Unlike neural network estimation of the dynamics of speech motor control, via point estimation was essentially spline fitting and, using ever-faster workstations, the time-frame for analysis was again in the realm of seconds rather than hours and days.

Simultaneous with the use of such complex computational schemes, movement transduction and data acquisition hardware improvements allowed more dimensions of motion to be recorded. It became possible to record 3D motions (e.g., of the face and jaw) and many channels of EMG. Analyses comparing a few visually derived, articulator position measured in different phonetic environments no longer made sense. For one thing, there were now too many channels of behavior; it was very difficult for the user to visualize time-series displays of two motion dimensions for 7-8 flesh-point measures. Also, the increase in data channels encouraged more comprehensive analyses aimed at modeling the spatiotemporal behavior in real speech, rather than attempting to characterize between carefully contrasted phonetic differences [6, 24]. Thus, the focus of the development effort again turned to reducing the complexity of the represented behavior, this time through animation. For example, Vatikiotis-Bateson and Ostry [22] examined the rigid body motion of the jaw through

decomposition of its three dimensional motion into its component rotations and translations. While it is true that the published analysis was based on measures made from multi-channel time-series displays and 2D plots of one component against another, it was almost impossible to explain the relation between changes in the six-dimensional representation to the more conventional representation in Cartesian coordinates. Figure 6 shows a sample of the multiple display, user interface for a graphics animation package [5] that demonstrates the effects of the different rigid body components on the motion of a digitized jaw+skull.

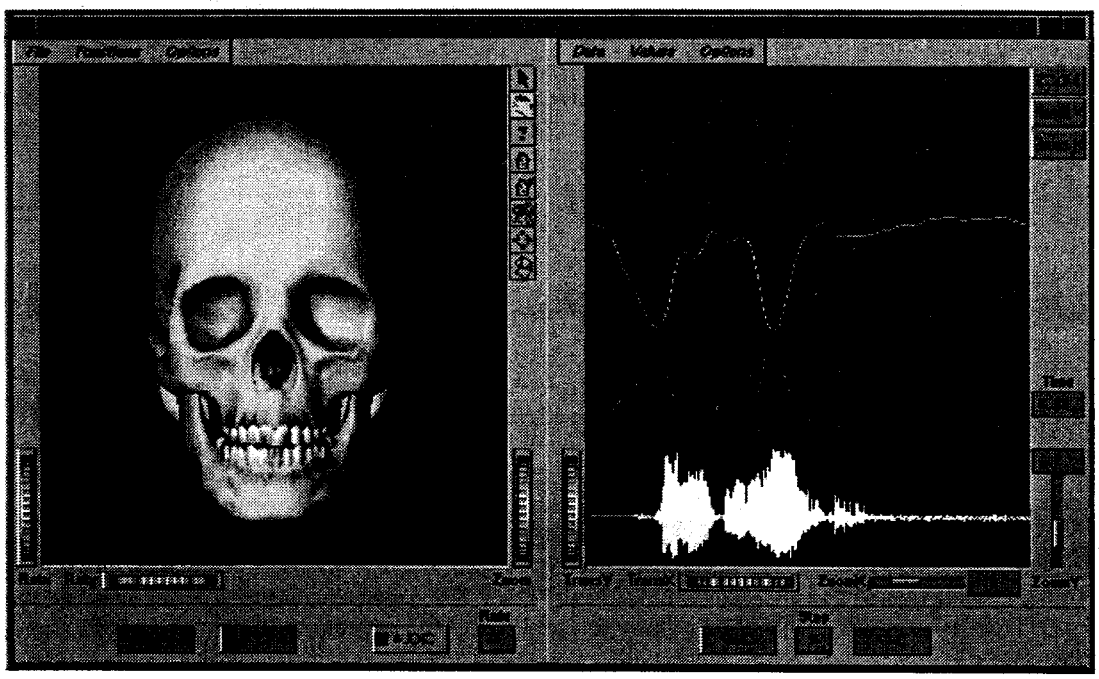

Fig. 6. Animated display windows for 3D jaw motion and time series of the six independent rigid body components using JawViewer.

The ICP. The graphics package shown in Figure 6 exemplifies the extraordinary (in speech research, at least) situation of the ICP where issues in speech science are pursued by a rich mixture of engineers, phoneticians, psychologists and their students. Similar to Haskins, the large group of researchers at the ICP have been loosely pressured by funding considerations into devising a coherent scheme for examining speech production behavior. In particular, their recent project goal has been to construct a "speech robot" that generates audiovisual speech output from phonemic text input[2].

The number of researchers and students involved with this project is large and their progress has been impressive. Similar to the early stages of modeling at ATR, the approach relies primarily on finding engineering solutions to the problems of controlling the output behavior — e.g., externally visible lip motion and the acoustics; relatively little attention has been paid to either the underlying muscle activity or, until recently, the behavior of vocal tract articulators. Instead, analogs to speech motor control have been used to generate articulator motions and statistical models of the articulator-to-acoustic mapping [11] have been used to estimate the output acoustics.

Among other things, verification of such an approach still relies heavily on simulations of the output behavior that can be judged perceptually. The model cannot, and should not, be held accountable for its biological plausibility, only for its ability to generate appropriate output. In order to do that, sophisticated animated simulations are required, particularly in the case of audiovisual behavior where lip and facial motions are being synthesized. Similar to Haskins programs, such as ASY, the software is generally fully interactive at every level allowing graphical manipulation of the model control parameters and quick synthesis of the intended output, be it a sequence of vocal tract configurations or audiovisual speech. In the presentation, samples of the various animations used by ICP in modeling speech production will be shown.

LOOKING AHEAD

As implied by the brief historical sketch of research in our own laboratories, there is a strong co-dependence between research issues and the tools designed to address them. Furthermore, the de-

gree of intensive hands-on experience with the manipulation and analysis of data affects the nature and direction of our conceptual development. In this section, we discuss the impact on speech production research of two developments that have been made possible by the recent, rapid improvements in hardware cost performance. They are (1) the transition from special purpose research tools developed in-house and at great cost to relatively inexpensive, general purpose data manipulation and analysis packages, and (2) the emergence of the "demo" as a major, and in some cases, the only means of communicating our research activities to others. As with any major development, these tools have their strengths and weaknesses, but are already deemed indispensible to those who just a few years ago relied on hand measurement and scatterplots for analysis and dissemination of their research results.

General Research Tools. A major obstacle in devising appropriate display and analysis tools in speech production in the past was the lack of commercial packages appropriate to our needs. This necessitated expensive and time-consuming development of special purpose tools like those described in the preceding sections. These tools in turn limited our research activity to the capability of the program. Also, collaboration between laboratories was much more difficult than it is today due to the dependency on specific computer environment. For two laboratories to collaborate required compatible computer platforms, which implied a commitment to long-term cooperation. Although not perfect, one of the great boons in the past few years has been the emergence of general purpose analysis and display packages. Compared to hardware development, these programs are cheap and exploit the increased speed and power of affordable computers while affording code-porting across a wide range of hardware platforms. One of the most popular examples of this is MATLAB (The MathWorks, Inc.).

Much of the current modeling work in our three laboratories is being done in MATLAB. The most tangible benefit is the tremendous savings in cost. Analysis routines and data can now be easily shared among collaborators without requiring expensive changes to anyone's machine environment. These large engineering packages contain well-implemented routines for signal processing, analysis, and display, and there is a well-conceived user interface for those cases where we need to write our own routines.

Use of such packages has already had and will continue to have a number of effects that can be seen as both good and bad. One is that there will be greater similarity of analysis across research domains. This is good in that the greater objectivity afforded by shared research paradigms should lead to better mutual understanding among researchers. Also, specific lines of inquiry can be pursued farther and faster when more people and points of view are involved. An example of such collaboration is the set of MATLAB routines for Sine Wave Synthesis (SWS) [14] provided by DAn Ellis on the Haskins web page.[3] Potentially less good will be the constant temptation for people to follow the herd by getting involved with the latest development at the expense of pursuing tough and possibly more interesting problems independently. Also, putting greater analytic power in the hands of non-specialists runs the danger of misuse. That is, we do not have to be signal processing engineers or mathematicians to take advantage of the many procedures available for our analyses. Indeed, our conceptual understanding may be only rudimentary or we may choose an analysis simply because others are using it. In our experience, however, such abuses tend to be short-lived in speech science. The heterogeneity of the speech research community usually results in the rapid ferretting out of such errors and increased use of common, well-documented procedures will aid this process. Moreover, making powerful analytic tools available to the masses has the great potential benefit of exposing problems to a wider range of insights (accidental or otherwise) than can be provided by the specialists alone.

The Almighty Demo. Discerning how best to communicate one's research activities to colleagues and non-specialists such as investors has always been a major task. Journal publications have become increasingly unsatisfactory for making our work understood to others, which coincides with the rapid increase in the number of professional meetings over the past several years. That is, more interactive venues that offer a wider range of visual aids now play a vital role in professional communication.

In the specific case of speech production research, there has been a shift in focus, as illustrated above, from descriptive comparisons based on a few measures elicited within a paradigm of phonetic contrasts to computationally intensive models that predict the observable output behavior across a range of realistic speech conditions. Even if phoneme-specific parametrization is used to control the model — e.g., visemes [2], via points [23] — the parameters themselves are usually in the form of eigenvalues or network weights, and are not very informative. Rather, the focus is on how well a particular model simulates the desired aspect of the speech behavior. For formal testing, the model output may be subjected to rigorous perceptual testing by means of an animated stimulus tape [3].

That same tape can also be used very effectively to demonstrate the outcome of the research ef-

fort to colleagues and non-specialists alike. For one thing, we are gratified that speech behavior, which occurs in time, can be animated appropriately. The very fact that something moves on the screen conveys the sense that significant results have been achieved. Also, audiences are easily impressed by the technical prowess of a good demo; textured surfaces rather than line drawings, 3D graphics, and good editing can easily blur the lines between the form and the content of the presentation for all but the most discriminating and skeptical audience. Our worry is that it is not just the audience that is susceptible to the seduction of a good demo. A good response reinforces our efforts to generate effective stimuli, which can become a time-consuming occupation quite independent of the research and may even dictate the direction of the research to best fit the presentation media.

As with any sudden change, we are confident that the negative aspects of this change of focus are primarily reactionary and therefore temporary. Recent technological developments provide powerful tools that will lead to better understanding. Publishing houses are already accommodating these changes by implementing electronic forms of multimedia distribution. The ability to inexpensively disseminate our models and simulations using richer forms of visualization can now more adequately inform readers and potentially engage them in ways that were not previously possible. In our field, books such as *Progress in speech synthesis* [18] are being published with accompanying CD-ROMs, thus affording the reader audio and even video demonstrations to augment the textual presentation. Journals such as *Speech Communication* have taken this a step further by using multimedia facilities in the review process. If we can avoid the temptation of being "demo-driven" and rely on such facilities for improved visualization only when they are critical for describing the methodology and results of our research, we have great hope that the emerging technologies will enhance our collaborative and scientific efforts.

REFERENCES

1. C. Benoît, T. Lallouache, T. Mohamadi, et al., in Talking Machines: Theories, Models, and Designs, G. Bailly, C. Benoît and T. R. Sawallis, Eds. (North-Holland, Amsterdam, 1992).
2. C. Benoît, T. Guiard-Marigny, B. LeGoff, et al., in Speechreading by Humans and Machines, D.G. Stork & M.E. Hennecke, Eds., (Springer-Verlag, Berlin, 1996).
3. C.P. Browman & L. Goldstein, Phonology Yearbook, 3,219-252 (1986)
4. F.S. Cooper, P.C. Delattre, A.M. Liberman, et al., J. Acoust. Soc. Am., 24, 597-606 (1952).
5. T. Guiard-Marigny, D. J. Ostry, & C. Benoît, in Proc. ICPHS-13, 3, 222-225 (Stockholm, Sweden, 1995).
6. M. Hirayama, E. Vatikiotis-Bateson, & M. Kawato, IEICE Trans., E76-A, 1898-1910 (1993).
7. J. Hogden, P. Rubin, & E. Saltzman, Bulletin de la Communication Parlée, 3, 101-116 (1995).
8. J.A.S. Kelso, B. Tuller, E. Vatikiotis-Bateson, et al., J. Exp. Psych.: Hum. Perc. Perf., 10, 812-832 (1984).
9. J.A.S. Kelso, E. Vatikiotis-Bateson, E. L. Saltzman, et al., J. Acoust. Soc. Am., 77, 266-280 (1985).
10. S. Kiritani, H. Imagawa, T. Takahashi, et al. Ann. Bull. RILP, 16, 1-10 (1982).
11. S. Maeda. Sp. Comm., 1, 199-229 (1982).
12. P. Mermelstein, J. Acoust. Soc. Am., 53, 1070-1082 (1973).
13. R. McGowan. Sp. Comm., 14, 19-49 (1994).
14. R.E. Remez, P.E. Rubin, D.B. Pisoni, & T.D. Carrell, Science, 212, 947-950 (1981).
15. P. Rubin, T. Baer, & P. Mermelstein, J. Acoust. Soc. Am., 70, 321-328 (1981).
16. P. Rubin & E. Vatikiotis-Bateson, in Animal Acoustic Communication: Recent Technical Advances, S. L. Hopp, M. J. Owren and C. S. Evans, Eds. (Springer-Verlag, Heidelberg, in press).
17. E. L. Saltzman, in Generation and Modulation of Action Patterns, H. Heuer, C. Fromm, Eds. (Springer-Verlag, Berlin, 1986).
18. J.P.H. van Santen, J. Hirschberg, J. Olive, et al., Progress in Speech Synthesis (Springer, New York, 1996).
19. C. L. Smith, C. P. Browman, & R. S. McGowan, J. Acoust. Soc. Am., 84, S128 (1988).
20. M.K. Tiede & E. Vatikiotis-Bateson, Proc. ICSLP-94, 1, 45-58 (Yokohama, Japan, 1994).
21. E. Vatikiotis-Bateson & J. A. S. Kelso, J. Phon., 21, 231-265 (1993).
22. E. Vatikiotis-Bateson & D. J. Ostry, J. Phon. 23, 101-117 (1995).
23. E. Vatikiotis-Bateson, M. K. Tiede, Y. Wada, et al., Proc. ICSLP-94, 2, 631-634 (Yokohama, Japan, 1994).
24. E. Vatikiotis-Bateson & H. Yehia, in Proc. ASA & ASJ 3rd Joint Meeting, 811-816 (Honolulu, HI, 1996).
25. Y. Wada and M. Kawato, Biol. Cyber. 73, 3-13 (1995).
26. Y. Wada, Y. Koike, E. Vatikiotis-Bateson, et al., Biol. Cyber. 73, 15-25 (1995).
27. J. R. Westbury, J. Acoust. Soc. Am., 95, 2271-2273 (1994).

NOTES

[1] Also, Yale University School of Medicine, New Haven, CT.

[2] ESPRIT/BR project no. 6975 — *Speech Maps.*

[3] http://www.haskins.yale.edu/haskins/MISC/SWS/MATLAB/matlab.html.