TOWARDS A PHYSIOLOGICAL MODEL OF SPEECH PRODUCTION

R. Wilhelms-Tricarico and J. S. Perkell
M.I.T. - RLE 36-581, 50 Vassar Street, Cambridge MA 02139, U.S.A.

ABSTRACT

At the periphery, speech production is a biomechanical process that has acoustic effects. To help understand this process and its control, we propose a model of the vocal tract that is based on biomechanics and adaptable to the individual speaker's anatomy; we report about initial and planned efforts to realize such a model; and we outline a hierarchical, modular control structure that transforms a stream of articulatory and acoustic goals into physiological signals that drive the vocal tract model.

INTRODUCTION

Just as the legs are not primarily "designed" for ball dribbling in soccer, the body structures that participate in the physical execution of speech processes, such as the lungs, larynx, tongue, jaw, and lips, did not evolve primarily for speech production. Speech as a physical process is based on audible effects of the movements of a poorly-understood, complicated biomechanical structure, which performs many non-speech purposes and functions. As with control of the many different uses of the hands, the speech process may require a special control structure in the human brain that manages to make appropriate use of the intricate biomechanics of the speech production apparatus.

Physiologically based speech production research usually investigates the hypothesized control structures and attempts to explain observations and measurements of speech production processes, in terms of functional constructs such as compensatory articulation, coarticulation, motor equivalence and articulatory invariance. Unfortunately, the complexity of the biomechanical and neurological system that makes up the production apparatus has often been characterized by oversimplified models which make unsupported or unrealistic assumptions about the nature of articulatory dynamics and kinematics.

We argue that without taking biomechanics into account, many empirical observations of speech production can only be explained with ambiguity, because no objective criteria exist to establish a distinction between what are mechanical effects and what are effects of the control system. In order to overcome the difficulty of establishing more clearly the domain of the different system interactions that influence speech motor output, a physiologically- based vocal tract model will be built. The methods and the current state of the modeling are reported in this article. In addition, we continue a discussion about the structure of a controller, in order to be able to express current ideas about speech motor control in a computational model.

TASK DYNAMICS

The most advanced, comprehensive production model to date has been developed at Haskins Laboratories (see [14] for a review). The Haskins model forms the core of a "dynamical approach to gestural patterning in speech production", which attempts "to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract, discrete, context-independent units, with the empirical observation of continuous, context-dependent interleaving of articulatory movements" [18]. The fundamental invariant unit is the abstract gesture. Combinations of abstract gestures underlie phonetic segments.

In this approach, a "task dynamic" model is the controller for an articulatory synthesizer, which is a geometrical model in the midsagittal plane. Computations of area functions are based on the articulator configurations, in combination with formulae derived from static three-dimensional data. An utterance-specific "gestural score" provides the input to the task-dynamic model in the form of sequences of activation pulses for the abstract gestures, following Browman and Goldstein [3].

In the task dynamic model, the formation and release of linguistically-significant vocal-tract constrictions are specified in a "tract variable" coordinate system. Articulatory movement is generated by modeling the influence of each discrete abstract gesture in this coordinate system as a time-invariant linear second order system. Since all dynamical properties of this system reside in the controller, the biomechanical properties of the vocal tract are not represented explicitly. The model accounts for coarticulation (as "coproduction" of sequences of (partly) overlapping abstract gesture complexes), and overlapping influences of multiple abstract gestures and tract variables on movements of individual articulators (as a result of "blending" of abstract gestures).

Limitations and Improvements

Perhaps the most important limitation of the Task-Dynamic Model is the concentration of its dynamical properties entirely in the abstract task space, when it is likely that a significant proportion of the kinematics of speech are determined by the anatomical and biomechanical properties of the peripheral production mechanism. An inadequacy of such simple task-space dynamics has also been shown in characterizing arm movements [9]. Many types of actual movement appear to be characterized by more gradually-increasing accelerations, and depending on the movement objectives, movements may be programmed (in part) according to optimization principles such as minimization of expended effort (cf. [13],[12]). If the non-linear biomechanical (dynamical) properties of the vocal tract were included in the model of the physical plant, then physically-based movement optimization criteria could be explored, simulations should be more accurate, and the form of the actual underlying control might be investigated with more revealing results.

Another limitation, less fundamental than the first, is seen in the almost axiomatic assumption of linear second order dynamics in the task space. Originally, it was proposed that in the task dynamic model, the programs of inter-articulatory coordination are "task-specific autonomous (time invariant) dynamic systems that underlie an action's form as well as its stability properties" (See Salzman and Munhall, [18] p. 337). This sufficiently general definition can, in principle, cover a large class of dynamic systems that may be needed to describe the dynamics of underlying coordinative structures. According to task dynamics, the movement from one segment to the next can be understood as a transition from the influence of one dynamic regime to the next. However, in the further development of the model, the understanding of task dynamics has been reduced by some to thinking of *second order* dynamic systems as the only possible model, resulting in a tendency towards over-simplification in which "everything" is to be accomplished by point attractors and limit-cycles. This limitation can be overcome by considering more general control systems, which by themselves can be dynamical systems, namely motor pattern generators. For modeling motor synergies, motor pattern generators have been proposed previously to simulate reflex behavior in animals (see [10]).

The task dynamics model also assumes generally that the movement goals of the abstract gestures are defined in terms of vocal-tract constrictions. However, recent motor equivalence studies of the vowel /u/ indicate that its goal may be defined more appropriately in terms of the acoustic transfer function [17]. Other sounds may also have goals that are defined primarily in acoustic terms [6].

Finally, it has been suggested that the establishment of sound categories is influenced partly by anatomy [14, 16].

Further, individual morphological differences between speakers may have influences on the motor-planning of individual speakers. To investigate such hypotheses, the current restriction to two-dimensional geometric vocal tract models needs to be overcome.

A BIOMECHANICAL MODEL

Our first step towards a biomechanical vocal tract model is a three- dimensional model of the human tongue (cf. [20]). Compared to previous work on 3-D finite element tongue models, the current work in progress entails a more accurate description of motion, by using large-strain finite elements and accounting for inertia of the moving structures.

So far, a simplified tongue model has been implemented and tested; the model consists of 42 elements, and contains eight tongue muscles. Fig. 1 shows the shape of the fixed reference configuration of the model tongue and the muscle fiber directions of the styloglossus muscle. Because of the lack of an accurate model of tongue tissue, a pragmatic phenomenological muscle model was adopted. The stress in the muscle tissue has an active and passive component. The passive stress is modeled as a nonlinear- elastic, linear-viscoelastic response of the tissue to deformation. The active component is computed by a stress production model that takes into account the elongation and the rate of elongation or shortening of the muscle fibers.

Computational methods

The application of the finite element method to the discretization of the equations of motion results in a system of differential equations which have to be solved. The system of equations relates the forces, displacements and accelerations at each node. The complete system has the following form:

$$M\ddot{u} + J(u, \dot{u}, \Pi) = B + T(u, \dot{u}) \quad (1)$$

In Eq. 1 the global node displacement vector u contains the displacement vectors of each node. The internal force vector J depends in a nonlinear manner on the global node displacement vector u, the global node velocity vector $\dot{u}$, and further, upon a multi-tuple of parameters Π which influence the constitutive behavior of the matter (strain–stress relation). The parameters Π are activation levels of the muscles in the model. Virtual forces are computed to maintain the incompressibility of the tissue, which is essentially a geometric constraint on the movement. This is done by computing a pressure field that varies over time but is spatially constant or varies linearly within each element. The pressure field is contained in the internal forces in equation (1). The right hand side of the equation is the system of external forces. Forces such as gravity, which act upon the whole body, are included in the vector B, and surface forces are represented by $T(u, \dot{u})$. The surface forces are responsible for constraining the model's movements geometrically. They include the forces resulting from intra-oral air pressure during closures and those forces acting on the tongue when it contacts and slides along surfaces such as the hard palate. In the current model surface forces have not yet been implemented. The details of the derivation and implementation of the computational methods are described in [20].

Since the time-dependent muscular activation levels modify the constitutive equations of the muscle tissue, they influence the stress field in the continuum, which is computed based on the instantaneous strain and rate of strain in the tissue. The computed stresses give rise to node forces. Thus, the node forces are a function of the deformation and of the muscle activation levels. The varying muscle activity levels constitute a multidimensional parametric control of the system.

The model has been used mainly to achieve an operational state of the computer code and to show the feasibility of the proposed methods, in that some typical movements of the tongue could be realized in simulation experiments [20].

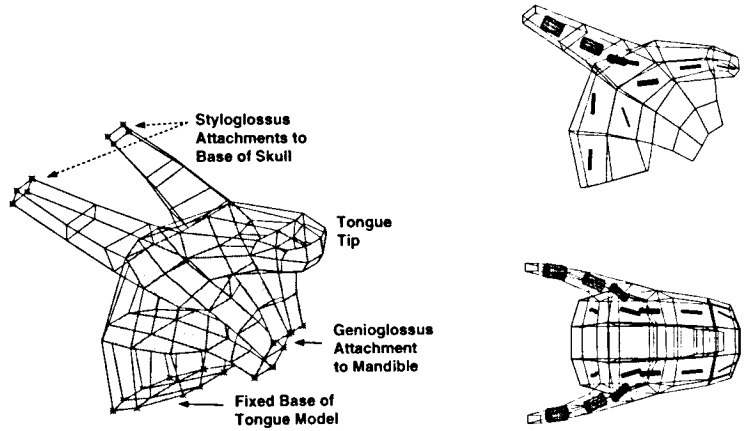

Figure 1: **Left:** *The tongue reference configuration with the initial placements of nodes. The nodes indicated with stars are fixed in the current implementation. They represent the connection of styloglossus muscles with the skull, the genioglossus with the mandible, and the base of the tongue in the plane of the hyoid bone. In the computational simulations, deformations of the tongue are computed as displacements of the nodes relative to their position in the fixed reference configuration.* **Right:** *Fiber directions of the styloglossus in lateral projection (above) and top view (below). Styloglossus fibers are specified on both sides in lateral elements. The long axes of the small cylinders represent the directions of the stress production by muscle fibers. The cylinders' diameter represent the relative amounts of maximal generated stress and can be interpreted as fiber density in the elements.*

CONTROLLER SCHEMES

Even at the current stage of development of the vocal tract model, it is possible to think about the general structure of a control mechanism that will compute the muscular activation time functions to steer the biomechanical model. This structure is not understood in terms of actual neurological functionality but rather as "software" which should simulate plausible functional components of the neurological controller. The controller transforms input signals that are described in terms of desired acoustical and/or articulatory goals into muscular activations which cause movement of the biomechanical plant.

Evidence for hierarchical organization

Perturbation experiments (cf. [2]) have provided evidence for the existence of mechanisms in speech production that use orosensory and internal feedback to control synergistic actions of multiple articulators for achieving functionally-specific goals. For example, if the lower-lip is unexpectedly impeded in its upward movement toward a bilabial closure for a /p/, the upper lip may move further downward than planned, with increased velocity to complete the closure (cf. [1]).

For articulators that are *not* biomechanically coupled, it has been shown [11] that the laryngeal articulation could be influenced by perturbing lip movements, further supporting the idea

that during speech production "coordinative structures" are programmed, which temporarily link articulators to achieve task specific goals. Weismer [19] describes this concept as follows: "Various articulatory gestures may be transiently linked to accomplish an articulatory goal, then unlinked as this goal expires and the next one arises." Such linking of articulators may involve the online use of both internal feedback and peripheral sensory information in the form of a "sensory template" that is specific for a motor control task. A sensory template is defined, according to Burgess [4] as "a central representation of the sensory receptor discharge that would be expected to occur during a movement if the movement is executed according to the plan".

Evidence of articulatory synergisms, *i.e.*, coordinative structures, also comes from motor equivalence experiments, which involve observation of many repetitions of the same behavior without the use of external perturbations. The term "motor equivalence" refers to the finding that the same goal is reached in more than one way (cf. [5]). Theoretically, across multiple repetitions, there can be trading relations (complementary covariation) in the relative contributions of: (1) multiple muscles to the same movement, (2) multiple movements to the same acoustically-critical vocal-tract cross- sectional area and (3) two area-function constrictions to the same acoustic transfer function.

Hypothetic controller structure

The movements to achieve sequences of articulatory and acoustic goals may be controlled by a hierarchical system that reduces the number of controlled degrees of freedom at each successively higher level.

The purpose of the hierarchical, modular controller is to control: multiple constrictions to determine the aerodynamics and acoustics of the vocal tract, multiple articulator movements for each constriction, and multiple muscles for each articulatory movement (see [15]). This hierarchy can be expressed by making the assumption that the controller has three hypothetical levels: The lowest level incorporates control structures, which generate synergistic muscle actions that result in simple gestural movements, such as raising the tongue blade or rounding the lips. The next level orchestrates the "elemental gestures" of the lower level to perform articulatory tasks that can be best described as creating vocal tract constrictions with certain characteristics (manners), for example creating an appropriate constriction for a vowel or producing a bilabial stop closure or a dento-alveolar constriction for a fricative. The third level, which orchestrates both lower levels, receives input signals that are described in terms of both desired acoustical consequences of articulation and/or as articulatory goals directly. The selection of acoustic goals and articulatory goals at the highest level comprises the translation of a hypothetical symbolic representation of speech into control actions.

Jordan and Rumelhart's distal learning strategy will be used as a paradigm for the implementation of each level of the controller, starting at the lowest level. The psychological and neuro-physiological idea of the internal model, or efferent copy, appears in this strategy as a forward model. The tentative general structure, shown in Fig. 2, has two components. One component (the controller C) maps from "intentions" (i) to motor commands (u), and the other component, called the forward model (FM), from motor commands to predicted sensations ($\hat{s}$). The forward model is trained using the difference between predicted sensation and actual sensation (s), which arise in the plant (P) as the result of the controlled actions. The composite system (C and FM) is trained using the difference between the desired sensations (d) and the actual sensation. See Jordan and Rumelhart [8] for further details.

In this context, Fig. 2 is a sketch of the first level controller. Since the biomechanical plant (P) is a dynamic system, the internal model of the lowest control level will also be a dynamic system (but without neural transmission and biomechanical response delays). The plant transforms the motor control input u and its current state x into two types of sensory results, s, and

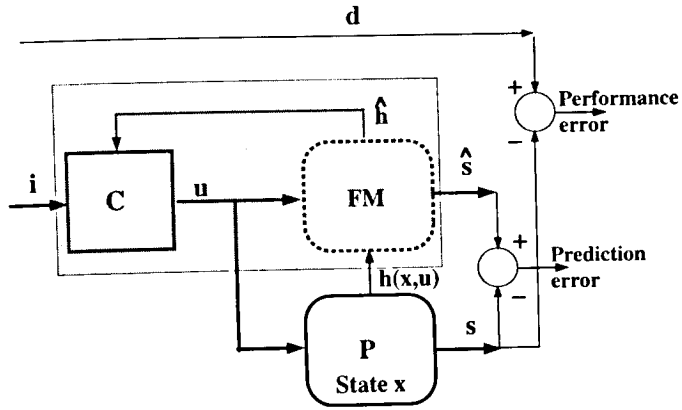

Figure 2: *The composite learning strategy (adopted and modified from Jordan, Flash and Arnon, 1993). The controller (C) transforms intentions (i) into motor commands (u). The forward model (FM) predicts $\hat{s}$ as the sensory result of the action u on the plant. The difference between the predicted sensation $\hat{s}$ and the actual sensation s, the prediction error, is used to optimize the forward model during training. The forward model is sensitive to the state of the plant; it receives input $h(x,u)$ that contains information about the state (for example muscle length information). Feedback delays of plant outputs are not shown.*

h(x,u) which makes the plant's state x (consisting of all displacements and velocities) partially observable by the forward model. The signal h(x,u) may contain measures of the length and rate of change of length of the muscles. The forward model learns to map motor commands (u) and current observables of the state (h(x,u)) into estimated sensory output ($\hat{s}$). It further learns to predict the development of relevant observables that are related to the plant state, shown as $\hat{h}$. Once the forward model is trained, the actual controller (C) relies on the estimated information ($\hat{h}$) about the state of the plant. This amounts to "internalizing" a feedback loop in the controller-forward model composite system. The purpose of the internal model is to "mimic" aspects of the plant, and to represent sufficient information to allow predicting the result of a control action on the plant. This information is represented in the forward model as a (learned) mapping from the current input u and the current state of the forward model $\bar{x}$ to the delayed sensory output $\hat{s}$ and the delayed estimated observables $\hat{h}$. The state $\bar{x}$ of the forward model is not necessarily an estimate of the actual state of the plant. For instance, if the forward model is implemented as a neural network structure, its state variables do not correspond to the state variables in the biomechanical model. However, the network learning should result in an input-output behavior that resembles the input-output behavior of the biomechanical model.

In view of the complexity of the biomechanical model and the speech motor control task, it will certainly be necessary to subdivide the overall control problem on each level into smaller ones. Subdivision on the lowest level is particularly sensible because the biomechanical plant consists of parts (e.g. the tongue body, tongue blade, mandible, lips, velum) that can act quasi-independently. Subdivision at the next level is motivated by the possibility for control of constrictions at different locations along the vocal tract with different articulators and manners. Another motivation for sub-

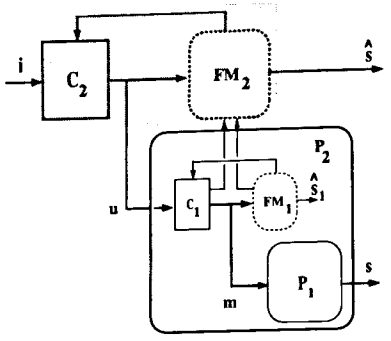

Figure 3: *The second level of the controller, consisting of the controller C_2, generates higher level motor commands u and controls the plant P_2 which is an encapsulated system consisting of the biomechanical model and the lowest level controller C_1. The biomechanical plant P_1 is driven by muscle activation functions m which are generated by the lower level controller C_1.*

division into subcontrollers for special tasks, in particular on the second level, is seen in recognized structures for interarticulatory coordination, such as between jaw and lips. Jordan and Jacobs [7] have extended their previously proposed competing experts paradigm to constitute a hierarchical architecture [the "hierarchical mixture of experts" (HME) paradigm] that allows such a subdivision of control problems.

By starting to build the controller from the bottom to the top, from simple movement models to complex movement models, each level of the controller is designed to achieve a reduction of complexity and degrees of freedom for the higher level controllers. For example, (sub-)controllers in the second level do not have to care about and operate without knowledge of individual muscle states, since that knowledge is incorporated into the lowest level of control. The internal model built into the lowest level of the control includes a partial (but sufficient) representation of the biomechanical model. The next higher level operates on an "encapsulated" lower level and its internal model includes a representation of the effects of combined generalized motor commands (such as "tongue tip raising" and "tongue back lowering") on sensory output, but only to the extent as it is relevant for the second level. Stated differently, the higher levels of the controller receive more general intentions, issue more global motor commands, and result in more abstract sensations. Fig. 3 sketches this arrangement for the two lower levels. The controller of the second level C_2 controls the augmented plant P_2 which consists of the lower level controller C_1 and the biomechanical model. Before controller C_2 can be trained properly, the forward model FM_2 of the second level needs to be trained to include a partial representation of the augmented system P_2, including the prediction of state-related information of the controller in P_2. The third level of the proposed hierarchy will operate on a plant formed by encapsulating the presented structure, and augmenting it further by adding another level of acoustical ouput resulting from computations in an extended biomechanical and acoustical model.

CONCLUSIONS

We have outlined a physiologically based speech production model and have reported the first steps taken in its implementation. Development of the biomechanical and control models will be coupled closely to the morphological (MRI), kinematic and acoustic data from individual speakers. A reasonably faithful model of the vocal tract biomechanics coupled with a pragmatically motivated, hierarchical and modular control structure, should permit investigations that allow greater insight into the actual underlying control strategies.

REFERENCES

[1] Abbs, J. H., Gracco, V. L., and Cole, K. J. (1984) Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming. *J. Motor Behavior*, 16(2):195–231.

[2] Abbs, J.H. and Gracco, V.L. (1984) Control of complex motor gestures: orofacial muscle responses to load pertubations of lip during speech. *Jounal of Neurophysiology*, 51:705–723.

[3] Browman, C.P. and Goldstein, L. (1992) Articulatory phonology: An overview. *Phonetica*, 49:155- 180.

[4] Burgess, P.R. (1992) Equilibrium points and sensory templates. *Behavioral and Brain Sciences*, 15(4). Commentary to Bizzi et al. (1992), Does the nervous system use equilibrium-point control to guide single and multiple joint movements?

[5] Hughes, O.M. and Abbs, J.H. (1976) Labial- mandibular coordination in the production of speech: Implications for the operation of motor equivalence. *Phonetica*, 33:199–121.

[6] Johnson, K., Ladefoged, P., and Lindau, M. (1993) Individual differences in vowel production. *J. Am. Soc. Acoust.*, 94(2):701–714.

[7] Jordan, M. I. and Jacobs, R. A. (1993) Hierarchical mixtures of experts and the EM algorithm. Computational Cognitive Tech. Rep. 9301, MIT.

[8] Jordan, M.I. and Rumelhart, D. E. (1992) Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354.

[9] Katayama, M., and Kawato, M. (1992) Visual trajectory and stiffness ellipse during multi-joint arm movement predicted by neural inverse models, *Technical Report*, ATR Laboratories, Kyoto, Japan.

[10] Liaw, J-S., Weerasuriya, A., and M.A., Arbib. (1994) Snapping: A paradigm for modeling coordination of motor synergies. *Neural Networks*, 7(6/7):1137–1152.

[11] Munhall, K.G., Löfqvist, A., and Kelso, J. A. S. (1994) Lip-larynx coordination in speech: Effects of mechanical pertubations to the lower lip. *J. Acoust. Soc. Am.*, 95(6):3605-3616.

[12] Munhall, K.G., Ostry, D.J., and Parush, A. (1985) Characteristics of velocity profiles of speech movements *J. Exp. Psych.*, 11:457–474.

[13] Nelson, W.L. (1983) Physical principles for economies of skilled movement, *Biological Cybernetics*, 46:135–147.

[14] Perkell, J.S. (1991) Models, theory and data in speech production. In *XIIth International Congress of Phonetic Sciences*, volume 1, pages 182–191, Aix-en-Provence, France.

[15] Perkell, J.S. (1994) Articulatory processes. To appear in J. Hardcastle and J. Laver (eds.), A Handbook of Phonetic Science.

[16] Perkell, J.S. (1995) Properties of the tongue help to define vowel categories: Hypotheses based on physiologically-oriented modeling. *Journal of Phonetics.* in press.

[17] Perkell, J.S., Matthies, M.L., Svirsky, M.A., and Jordan, M.I. (1992; in press) Goal-based speech motor control: A theoretical framework and some preliminary data. J. Phonetics.

[18] Saltzman, E.L. and Munhall, K.G. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382.

[19] Weismer, G. (1988) Speech production. In *Handbook of Speech-Language Pathology and Audiology*, chapter 8. B.C. Decker Inc.

[20] Wilhelms-Tricarico, R. (1995) Physiological modeling of speech production: methods for modeling of soft-tissue articulators. *J. Acoust. Soc. Am.*, in press.