VOWEL-VOWEL PLANNING IN ACOUSTIC AND MUSCLE SPACE

M. Zandipour [1,2], F. Guenther [2,1], J. Perkell [1,2], P. Perrier [3], Y. Payan [4], P. Badin [3]

[1] Research Lab of Electronics, M.I.T., Cambridge, MA, USA
[2] Dept. of Cognitive and Neural Systems, Boston University, Boston, MA, USA
[3] Institut de la Communication Parlée, France
[4] Techniques de l'Imagerie, de la Modélisation et de la Cognition, France
majidz@speech.mit.edu

ABSTRACT
The objective of this research is to investigate the planning and control of vowel-to-vowel sequences using a computer model of the vocal tract. The vocal tract model, which is based on a French male speaker, consists of a biomechanical tongue model (Perrier et al., 2003) with the addition of jaw (rotation and translation) and lip (opening and protrusion) movements. A comparison of acoustic data from the subject on whom the model is based to simulation results from the model for /i/-/a/, /i/-/e/, and /i/-/u/ movements is presented. Results show that planning a V-V sequence in either acoustic space (F1, F2, and F3) or motor space (in terms of lambda values according to the equilibrium point hypothesis, EPH) produces formant trajectories similar to those of the subject's data, and at the same time has a smooth progression of muscle lengths between the two vowels. However, the modeling does not account for all aspects of the data.

INTRODUCTION
Speech production, like many other skilled movement tasks, is learned and refined gradually. The integration of sensory information and efference copies of motor commands is likely brought about by learned inverse and forward models. An inverse model is needed to provide an accurate estimation of the motor commands required to achieve a desired sensory response, whereas a forward model predicts the sensory consequence of an action and minimizes the delay and noise between the sensory inputs and motor commands (Wolpert, et al., 2001).

The top panel in Fig. 1 schematizes this general concept of motor control adapted to speech production, based on the assumption that speech movements are planned in an acoustic/ auditory reference space (see also: Perkell, et al., 2000). A comparator measures the difference between the estimated current position and the desired target in acoustic/auditory space. This information is then sent to an adaptive inverse model responsible for generating the correct motor command. The forward models in Fig. 1 are state estimators based on sensory feedback and internally generated predictions of articulatory position (efference copy). The forward models transform vocal tract configurations (in muscle space) into corresponding acoustic parameters: a many-to-one mapping. Learning these mappings depends on acoustic and orosensory feedback, and on the efference copy of the motor commands. An alternative view is schematized in the bottom panel, in which a plan of speech motor control is based on transformations in muscle (motor) space. A comparator measures the difference between the current estimated and the desired target muscle lengths. This information is then used by an adaptive inverse model to generate the correct motor commands. This approach also employs both forward and inverse models to carry out the plan (Kawato, 1989).

Perkell et al. (2000) proposed three hierarchical levels for planning speech production, based upon solving the inverse kinematics and the inverse dynamics of the speech production system.
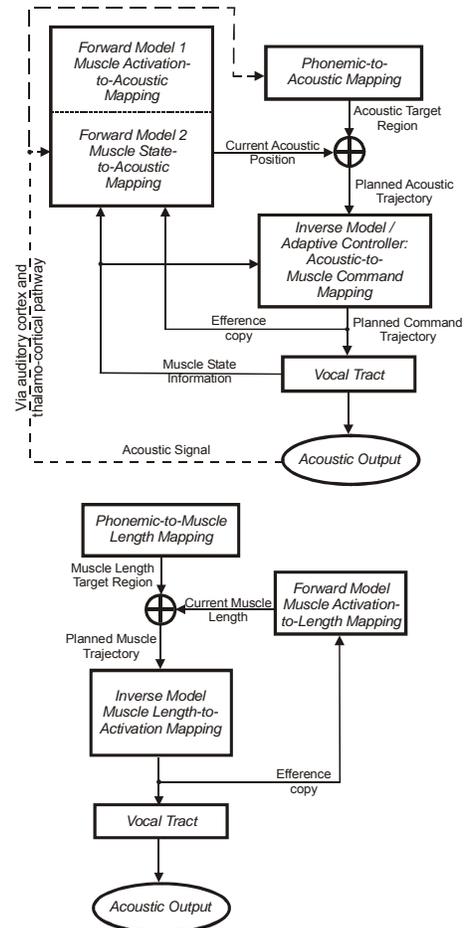
Several competing solutions to the inverse dynamic problem have been proposed in the arm movement control literature; two of them are considered here. According to the equilibrium point hypothesis, EPH (Feldman, 1986), the brain does not explicitly compute the necessary forces, but instead, controls the movement by shifting (in time) an "equilibrium point" at which all muscles and external forces are balanced, utilizing the spring-like properties of muscles and feedback loops. Another hypothesis postulates that the brain creates and adaptively updates internal models of the body dynamics and the environment (Wolpert, et al., 1998).

Computational models can be used to simulate and explore the implications of different control hypotheses, which can then be compared with experimental observations. The main objective of this research is to explore the coordinate space in which speech movements are planned. Using a model of the vocal tract based on a two-dimensional biomechanical tongue model (Perrier et al., 2003), the formant (F1, F2, and F3) trajectories of vowel-vowel (V-V) sequences are generated and then compared to data from one subject. *If the planning in acoustic space results in a motor space trajectory comparable to that of the subject's data (simulated by the subject's vocal tract model), then speech movement planning may occur in acoustic space. On the other hand, if planning in motor space results in a formant trajectory like that in the subject's data, then speech movement planning may occur in motor space. Alternatively, if both methods produce V-V trajectories similar to the subject's data and exhibit a smooth progression of muscle lengths, then neither planning hypothesies can be rejected.*

Figure 1: Schematic of speech production based in acoustic (top) and motoric space (bottom panel).
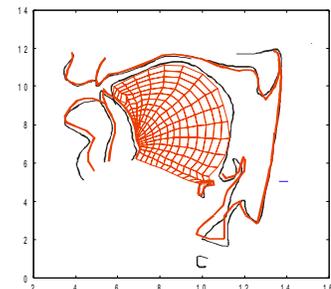
METHODS

The Model

The biomechanical tongue model (Perrier, et al., 2003) has 221 nodes (17 x 13) defining 192 quadrilateral elements in the midsagittal view (Fig. 2). Seven muscle synergies were modeled within this finite element (FE) mesh. Models of jaw opening (rotation and translation), lip opening, and lip protrusion were added to the biomechanical tongue model in order to simulate utterances in a more natural way. To identify vowel-specific values of λ (the static threshold length for force generation) for each tongue muscle, x-ray images for each of the vowels /i/, /a/, /u/, /e/, and /o/ at steady state were compared to their simulation of the vocal tract at steady-state (Fig. 2). Electromyography (EMG) data presented by Alfonso et al. (1982) from a speaker of

Figure 2: Vocal tract shapes from x-ray tracing (black) and the model (red) for /i/.

American English were used to determine the muscle activation for each vowel. To obtain the acoustic transfer function of the shape of the vocal tract, a model described by Beautemps et al. (1995) was employed, and its parameters were optimized such that the values of formants (F1, F2, and F3) were within ±10% (max error) of those from the subject's data.

Fig. 1 schematizes two forward models: one learns the muscle-state-to-acoustic mapping, and the other one learns the mapping between the efference copy of the motor command and the acoustics. Two neural networks based on Hyperplane Radial Basis Functions (HRBF) were trained to acquire these models (see also Guenther et. al 1998). The network is presented with a data set containing training pairs, each composed of a vector from the muscle-length/motor command space and a desired acoustic target (teaching signal). Each network is composed of 10 input nodes, representing the 10 dimensions of muscle-length/motor command space, 162 nodes at the middle (hidden) layer, and 3 output nodes, representing the 3 formants (F1, F2, and F3). Both networks converged after 500 iterations. The performance of each network was measured by the difference between the actual and obtained values of the formants from a different set of simulation results.

The theoretical detail of directional mapping from the auditory/acoustic space to the motor command space, i.e. an inverse model, has been discussed in Guenther, et al. (1998). In this study the inverse model is based on the *mixture of experts* of the *n* nearest neighbors, in which each nearest neighbor expert module learns a localized inverse model (IM) by employing the reinforcement learning method. The outputs of IMs are weighted based on the inverse of their distances in the muscle-length space; i.e., the further each module in the muscle space is from the current position, the less contribution to the total output. The final output is the sum of the normalized weighted output from each module (Schaal and Atkeson, 1998). In addition to smooth transitions and generalization from one IM to another, the multiple IM system has the ability to learn novel mappings, while maintaining the past. Each IM expert is an HRBF neural network consisting of 3 input nodes (3 formant trajectories, ΔF), 10 output nodes (10 motor-command trajectories, $\Delta\lambda$), and 8 nodes at the hidden layer. A data set containing 500 [$\Delta F\ \Delta\lambda$] pairs was collected by randomly perturbing the motor-command λ's associated with each muscle by 1/400 of its total range, and then obtaining the corresponding formants from the Forward Model of the vocal tract. Because the objective of the IM was to learn the formant-trajectory-to-λ-trajectory mapping, and because formants-to-λ is a many-to-one mapping, the error was measured by the difference between the desired and approximated formant trajectory vectors. The maximum error was set at 2%. All 1907 expert networks converged by the 200th iteration.

According to the proposed theories of speech motor control, planning a speech sequence (concatenation of phonemes) can be accomplished in one of the following ways:
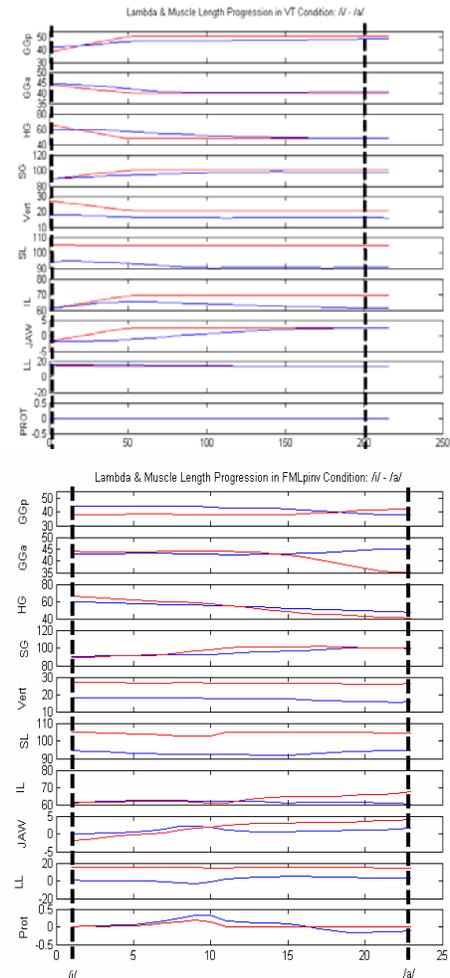
1. In *Motor (muscle) space*: By defining the target muscle configurations, and then interpolating between the present and desired configurations. This scheme works if and only if the system has <u>learned</u> the mapping between motor and acoustic spaces of each phoneme by realizing the acoustic consequence of each motor command. To acquire the acoustic outcome of this planning scheme, the initial and final configurations (in λ's) of each vowel target were entered into the vocal tract model, and then the formants were calculated from the derived area functions of the vocal tract.

2. In *Acoustic (formant) space*: By defining target acoustic settings, and then interpolating linearly between the present and desired configurations. Because the mapping from acoustic-space to motor-space is one-to-many, finding the correct motor command requires calculation of an inverse of the Jacobian that maps motor commands to acoustic perceptual variables. There are two ways of approximating the inverse of the Jacobian: a) by direct calculation of the Moore-Penrose pseudoinverse of a matrix; or b) by using an artificial neural network to approximate the pseudoinverse.

Simulations

A set of /i/-V sequences was simulated using the model described above, where V was /a/, /e/, or /u/, and the results were compared to data from the subject. The transition time between the two vowels was set to 200 milliseconds, comparable to the subject's data. The model produces an acoustic trajectory, muscle lengths, and muscle activations, which are presented below. The results presented below are from simulations with the vocal tract model when planning a V-V sequence in motor space (1) or in acoustic space (2a,b), such that:

1. Vocal Tract only, the beginning and ending vowels' λ's were interpolated according to the method of the λ-model (**VT** condition; $t_{V\text{-}V}$ = 200 ms).

2a. A forward model was trained to learn the mapping between the λ values and formants (Forward Model 1 in Fig. 1, top panel). It was then used to calculate the inverse of the Jacobian algorithmically which approximates the motor trajectory while planning in acoustic space (*FMpinv* condition: direct calculation of pseudo-inverse of the Jacobian using Forward Model). This method is based on planning a movement in acoustic space, in which at any point in time, a desired vector in formant space (ΔF_d) (whose length is 130 Hz - a value determined in an *ad hoc* manner) and points toward an acoustic target. Hence, theoretically, an acoustic trajectory should be a straight line in acoustic space because of linear interpolation in that space.

Figure 3: Blue lines: muscle lengths, jaw opening, lip opening, and lip protrusion vs. time for an /i-a/ sequence, when simulating in *VT* condition (motor space planning; top panel), and in *FMLpinv* condition (acoustic space planning; bottom panel). Red lines: represent progression of the λ target (motor command). In the bottom panel, the X-axis represents the number of steps taken by the pseudo-inverse algorithm to interpolate an /i/-/a/.

2b. A forward model was trained to learn the mapping between the λ's and formants. A motor-command trajectory (change in λ) while planning in acoustic space was then approximated using a network of *Mixture of Experts* based on HRBF as discussed above (*FMLpinv*

condition: Learned pseudo-inverse of the Jacobian using Forward Model). In this condition the **ΔF** to **Δλ** mapping is learned without any explicit knowledge of any Jacobian matrix, whereas in *FMpinv* condition many Jacobian matrices (**ΔF** to **Δλ**) and their inverses (**ΔF** to **Δλ**) along each formant trajectory are calculated.

RESULTS

Results of V-V planning for /i/-V were similar. Due to page limitations, only the results of the /i/-/a/ simulation is presented. Figure 3 show the progression of the λ targets (the motor commands; red lines), and muscle lengths (blue lines) in time for an /i-a/ sequence, when simulating in the *VT* (motor space planning; top panel) and *FMLpinv* (acoustic space planning; bottom panel) conditions. Note that due to interacting forces from different tongue muscles on each node of the FE mesh, the muscle lengths do not reach their intended targets. In addition, the transition of each muscle's length is a smooth curve, which is not necessarily monotonic due to interacting forces; e.g. GGa during the /i/. Planning in acoustic space (bottom panel) resulted in a smooth evolution of the λ values, though not a ramp like that of the λ model. Also, the λ progression is mostly monotonic, and any slight non-monotonicity can be explained by the imperfect nature of the forward model; i.e. it is an <u>estimate</u> of the vocal tract model. Note that what is different between the λ's in *VT* and *FMLpinv* conditions is the shape of the ramp, and importantly, contrary to the *VT* condition, in *FMLpinv* the λ's for different muscles are not synchronized in time. In addition, the muscle lengths in *FMLpinv* condition progressed smoothly and non-monotonically, as shown in Fig. 3 (blue line).
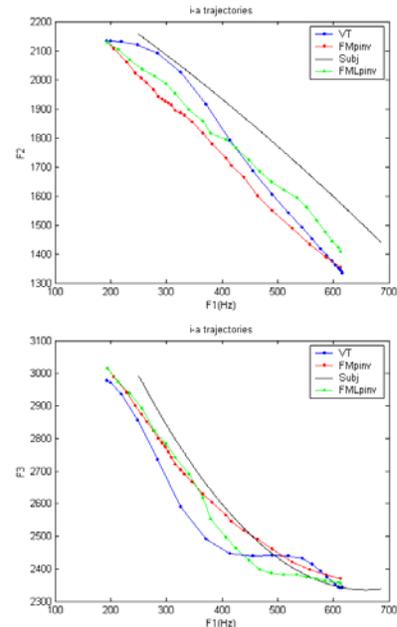
Figure 4: Formant trajectories of different planning scheme in F1-F2 (top panel) and in F1-F3

Fig. 4 depicts the comparison of the acoustic trajectories of /i/-/a/ for the subject's production data and the *FMLpinv* condition, in which the **Δλ** was approximated from the desired **ΔF** using a network of *Mixture of Experts*. The results from the *FMpinv* and the *VT* conditions were also added to show: a) how the learned inverse model was able to capture effectively the pseudoinverse of the Jacobian; b) how the learned forward and inverse model combination can produce results similar to the subject; and c) how close the results of these various conditions are to each other and to the subject's data despite different planning strategies.

Carré et al. (2001) showed that the shape of the trajectory between the two vowel targets was perceptually important; however, their data demonstrated a wide range of perceptual tolerance of variation in interpolations, synchronization, and transition duration. Hence, it is reasonable to conclude that the V-V trajectories in Fig. 4 likely would be perceptually indistinguishable.

SUMMARY AND DISCUSSION

It is hypothesized that in speech motor planning, articulator movement is planned within an auditory frame of reference to achieve an acoustic goal, or alternatively, planned within a motor frame of reference to reach a motor goal corresponding to its acoustic target. The results of the

present study (considering caveats about the vocal tract model) do not reject either of the two opposing hypotheses of speech movement planning, which rely on either acoustic or motor frames of reference. First, our results show that both schemes produce comparable results. Specifically, we have shown that planning a V-V sequence in motor space (as defined by the λ model) could produce formant trajectories similar to data. Second, we have also shown that planning a V-V sequence in acoustic space (as defined by formant frequencies) could result in smooth muscle length transitions between the two vowel targets, when using a vocal tract model based on a bio-mechanical tongue. If additional research supports the present results, a model of speech production adapted from one proposed by Hikosaka et al. (1999) can provide a unified alternative: learning and planning a speech sequence occurs in two parallel cortical systems, one using auditory coordinates and the other using motor coordinates.

REFERENCES

Alfanso, P. J., Honda, K., Baer, T., and Harris, K. (1982). Multi-channel study of tongue EMG during vowel production. *103rd Meeting of the Acoustical Society of America, Chicago, IL.*

Beautemps, D., Badin, P., and Laboissière, R. (1995). Deriving vocal-tract area function from midsagittal profiles and formants frequencies: A new model for vowels and fricative consosnants based on experimental data. *Speech Communication*, *16*, 27-47.

Carré, R., Ainsworth, W. A., Jospa, P., Maeda, S., and Pasdeloup, V. (2001). Perception of vowel-to-vowel transitions with different formant trajectories. *Phonetica*, *58*, 163-178.

Feldman, (1986). Feldman, A. G. (1986). Once more on the equilibrium point hypothesis (λ-model) for motor control. *J. Motor Behaviour, 18*, 17-54.

Guenther, F. H., Hampson, M., and Johnson, D., (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105(4)*, 611-633.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., Miyach, S., and Doya, K. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neuroscience (TINS), 22(10)*, 464-471.

Kawato, M. (1989). Motor theory of speech perception revisited from minimum-torque-change neural network model. *Proc. of 8th Symposium on Future Electron Devices*, *Oct.*, 141-150

Perkell, J. S., Guenther, F. G., Lane, H., Matthies, M. L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., and Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *J. of Phonetics*, *28*, 233-272.

Perrier, P., Payan, Y., Zandipour, M., and Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *J. of the Acoustical Society of America*, *114(3)*, 1582-1599.

Schaal, S., and Atkeson, C. G. (1998). Constructive incremental learning from only local information. *Neural Computation*, *10(8)*, 2047-2084.

Wolpert, D. M., Ghahramani, Z, and Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, *5(11)*, 487-494.

Wolpert, D. M., Miall, R. C., and Kawato, M. (1998). Internal Models in the cerebellum. *Trends in Cognitive Sciences*, *2*, 338-347.