Task dynamic and articulatory recovery of lip and velar approximations under model mismatch conditions

Richard S. McGowan and Mindy Lee
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695

(Received 18 June 1994; revised 16 May 1995; accepted 27 September 1995)

An algorithm for recovering task dynamics and speech articulator movements from speech acoustics was tested under various model mismatch conditions. There was evidence of articulatory compensation to recover tract-variable (constriction) trajectories in speech produced with a lip approximation under sufficiently constrained conditions. However, in more extensive studies of lip and velar approximations, the recovered tract-variable trajectories were also different from those of the data-producing utterance. This phenomenon can occur because the matching criterion in the analysis-by-synthesis procedure is an acoustic criterion and the correspondence between the tract-variable trajectories and the acoustic output is not exact. While there may be some tract-variable compensation to attain a good acoustic match, there is evidence of a correspondence between how well the tract-variable trajectories match and how well the formant frequencies match in particular instances. © *1996 Acoustical Society of America.*

PACS numbers: 43.72.Ct, 43.70.Jt

INTRODUCTION

Recent work on recovering task dynamics and articulation from speech acoustics is extended in this paper. In the past work, it was assumed that there was a perfect match between the task dynamics of the vocal tract that created the acoustic data and that which was used in the analysis-by-synthesis procedure for task dynamic or articulatory recovery (McGowan, 1993, 1994, 1995). In the present work, mismatches between these task-dynamic specifications were tested to find how important it is to have a model that is close to the data-producing model in the recovery process. Such tests are important for the recovery of human articulation, because, unless a very large number of parameters are used to fit the articulatory model to the human vocal tract, there will be some mismatch between the model and the human tract.

In the present work, the data-producing model and the model used in the analysis-by-synthesis scheme were purposely mismatched. These mismatches were in the task-dynamic specification, particularly, the constraints on tract variable, or gestural, natural frequencies, and articulator weights. There were no mismatches in the anatomy and geometry of the vocal tracts, as the same articulatory synthesizers were used to produce the data and to perform the analysis-by-synthesis.

I. BACKGROUND

A. Task dynamics

Task dynamics provides a means of recruiting various articulators in the performance of constriction tasks during speech. The articulators refer to those of ASY, the Haskins articulatory synthesizer (Mermelstein, 1973; Rubin *et al.*, 1981). The articulators of concern are shown in Fig. 1(a), and they are jaw angle, JA; tongue body center vector angle and length, CA and CL; upper lip vertical position, ULV, and lower lip vertical position, LLV; and lip horizontal position, LH, which applies to both lips. In the task-dynamic model, the dynamics of tract variables, which are places and degrees of constriction, are each given a linear, second-order dynamics (Saltzman and Munhall, 1989). The tract variables of interest here come in two constriction location–degree pairs [Fig. 1(b)]. These are tongue body constriction location, TBCL, and tongue body constriction degree, TBCD; and lip protrusion, LP, and lip aperture, LA. If $\mathbf{z}$ is a vector of tract variables, then the set of dynamical equations can be written

$$\mathbf{M}\ddot{\mathbf{z}}+\mathbf{B}\dot{\mathbf{z}}+\mathbf{K}(\mathbf{z}-\mathbf{z}_0)=0, \tag{1}$$

where rest positions, or targets, are specified by vector $\mathbf{z}_0$, and spring constants, damping constants, and masses are specified by the diagonal matrices, **K**, **B**, and **M**, respectively. The interval of time for which a tract variable is given a target that is not the default position is called an *activation interval*. During an activation interval, a tract variable's damping ratio, natural frequency (equivalently, stiffness), and target are specified. The tract variables, $\mathbf{z}$, recruit the articulators, ϕ, of the articulatory model used in ASY. The transformations from articulators to tract variables are nonlinear and differentiable almost everywhere, and can be written formally as

$$\mathbf{z}=\mathbf{z}(\phi). \tag{2}$$

The Jacobian of this transformation is the matrix

$$\mathbf{J}=(J_{ij})=\left(\frac{\partial z_i}{\partial \phi_j}\right). \tag{3}$$

In the implementation used here, there were more articulators than tract variables, so that there were more columns than rows in **J**. The Jacobian can be used to write the dynamical equations [Eq. (1)] in the articulator space, rather than the tract variable space (Saltzman and Munhall, 1989):

$$\mathbf{M}(\mathbf{J}\ddot{\phi}+\dot{\mathbf{J}}\dot{\phi})+\mathbf{B}\mathbf{J}\dot{\phi}+\mathbf{K}(\mathbf{z}(\phi)-\mathbf{z}_0(\phi_0))=0. \tag{4}$$

 0001-4966/96/99(1)/595/14/$6.00

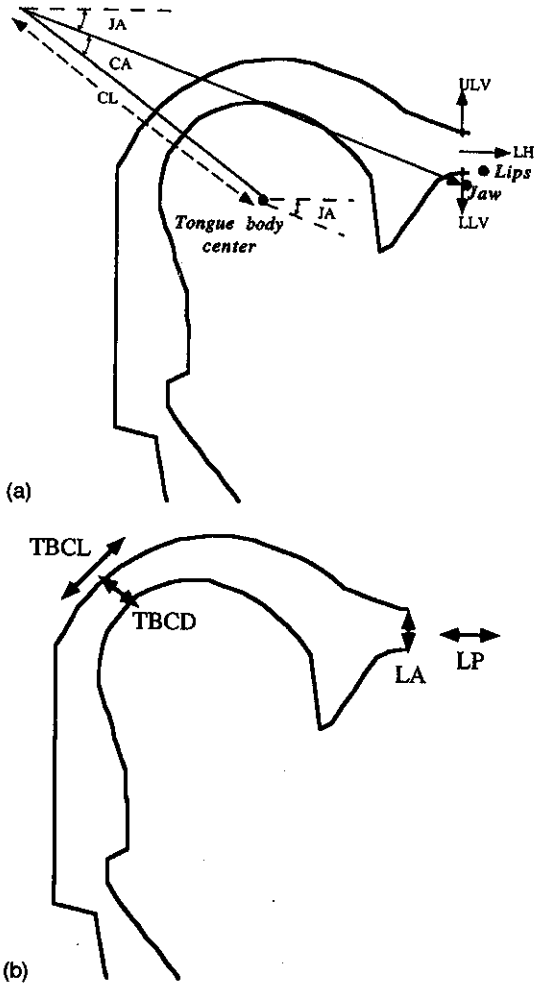

FIG. 1. (a) A midsaggital view of the vocal tract used in the Haskins articulatory synthesizer (ASY). (b) Tract variables in relation to the Haskins articulatory synthesizer vocal tract.

Writing the second-order derivative in terms of the lower-order derivatives for numerical solution gives

$$\ddot{\phi}=\mathbf{J}^*\{(\mathbf{M}^{-1}[-\mathbf{B}\mathbf{J}\dot{\phi}-\mathbf{K}\Delta\mathbf{z}(\phi)])-\dot{\mathbf{J}}\dot{\phi}\}, \tag{5}$$

where $\Delta\mathbf{z}=\mathbf{z}-\mathbf{z}_0$ and $\mathbf{J}^*$ is a weighted pseudoinverse. Assuming that the rows of $\mathbf{J}$ are linearly independent,

$$\mathbf{J}^*=\mathbf{W}^{-1}\mathbf{J}^T(\mathbf{J}\mathbf{W}^{-1}\mathbf{J}^T)^{-1}, \tag{6}$$

where $\mathbf{W}$ is the weighting matrix and the superscript T denotes transpose. In the present implementation, $\mathbf{W}$ is assumed to be diagonal, and hence $\mathbf{W}^{-1}$ is diagonal. The matrix $\mathbf{W}^{-1}$ multiplying $\mathbf{J}^T$ has the effect of multiplying the partial derivatives in row j of $\mathbf{J}^T$ with the same number, ${w_{jj}}^{-1}$. Thus the factor used to multiply partial derivatives was the same for a given articulator no matter which tract variable was involved. Note that the larger the weight w_{jj}, the smaller the weighted derivative of any tract variable with respect to the jth articulator. One can think of increasing the weight of the jth articulator as adding mass to that articulator, so that the "lighter" articulators are more likely to move to attain tract variable goals.

There were two types of mismatching allowed in the experiments reported here. One involved the articulator weights found in the $\mathbf{W}$ matrix, which were always fixed during recovery, and therefore not varied in the analysis-by-synthesis. In this mismatch condition, the data would be produced with one set of articulator weights and recovered with another set. The other mismatch in some experiments involved the constraints on the natural frequencies of the tract variables or, equivalently, the stiffnesses in the $\mathbf{K}$ matrix. (The natural frequencies of the tract variables are known here as gestural natural frequencies, or GNFs. As will be explained below, tract variables that constitute constriction location–degree pairs were constrained to have identical natural frequencies.) The gestural natural frequencies, the GNFs, were recovered variables, but they were constrained to be the inverse of the activation time interval. In a GNF constraint mismatch the data producing utterance had a GNF that did not meet this constraint.

B. Genetic algorithm

A genetic algorithm was used as an optimization algorithm to recover task-dynamic parameters, and hence tract variable and articulator trajectories, from the given acoustic data. The particular algorithm used in these computer experiments was a modified version of the simple genetic algorithm as described by Goldberg (1989, pp. 59–70). In the implementation of the genetic algorithm for the present experiments, acoustic data is given in the form of the first three formants are presumed to be given in 10-ms intervals for the length of the utterance. To initialize the algorithm, a population of chromosomes (100 in the present experiment) describing individual utterances are generated randomly. These chromosomes contain coded task-dynamic parameters that are necessary for running the model and which are allowed to be changed in this analysis-by-synthesis procedure. (Other task-dynamic parameters are not represented in the chromosomes, such as articulator weights, and they are fixed for a given run of the algorithm.) Parameters, such as the starting times and durations of activation intervals and targets of various tract variables, are binary coded using a Gray code (Forrest, 1993), and these are concatenated to form chromosomes. Before starting the iterative part of the algorithm, each chromosome is decoded, the specified utterance is synthesized in ASY, and the first three formants are tracked in 10-ms intervals, the same as for the data. A mean-square comparison is made between the output formants produced by the chromosome and those of the data, and that comparison is quantified as a fitness assigned to the chromosome:

$$\text{fitness}=\frac{1}{6N}\sum_{i=1}^{3}\sum_{j=1}^{N}\frac{(f_{ij}^{\text{data}}+f_{ij}^{\text{model}})^2}{(f_{ij}^{\text{data}}-f_{ij}^{\text{model}})^2}, \tag{7}$$

where i denotes formant number, j frame number, and N equals the number of 10-ms frames. Thus each individual is comprised of a chromosome containing a task-dynamic description and a fitness quantifying how well the speech produced by that description matches that of the data.

The iterative part of the algorithm consists of fitness proportionate selection, mating, and mutation. In fitness proportionate selection, an individual is selected with a probability proportional to its fitness for either mating or inclu-

TABLE I. Possible variable task-dynamic parameters during recovery with and without constraints applied.

Tract variable	Possible variable task-dynamic parameters without constraints applied		Possible variable task-dynamic parameters with constraints applied	
LA (lip aperture)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position target position	start activation time duration of activation (>100 ms) (GNF=inverse duration of activation. All motions are critically damped.)
LP (lip protrusion)	start activation time end activation time natural frequency	damping ratio target position articulator weights		
TBCL (tongue body constriction location)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position target position	start activation time duration of activation (>100 ms) (GNF=inverse duration of activation. All motions are critically damped.)
TBCD (tongue body constriction degree)	start activation time end activation time natural frequency	damping ratio target position articulator weights		

sion into the next generation without mating. The number of selections that are made in a given population is equal to the number of individuals. The least fit individual after selection, mating, and mutation was always replaced by the most fit individual of the previous generation in the elitist strategy implemented here. As each pair of individuals is chosen, a decision is made on whether they will mate. This decision is based on a weighted coin toss that had a probability of 0.8 for mating in these experiments. When mating occurs, the chromosomes of each individual is cut at the same randomly selected location on the chromosomes, and the resulting substrings are swapped between the individuals to form two children chromosomes. The individuals that do not mate are passed through to the next generation. Any bit in children or passed-through chromosomes had a 0.001 probability of mutation in these experiments. The children chromosomes replace the parent chromosomes in the next generation, and if the children are different from their parents, then they are assigned new fitnesses. To assign a fitness, a child chromosome is decoded into its task-dynamic parameter specification, speech is synthesized, formants extracted, and fitness evaluated based on Eq. (7). In the experiments here, this procedure was iterated for 60 generations, which was sufficient to obtain a very homogeneous population as measured by average Hamming distance from the fittest individual. This is a modified simple genetic algorithm, where the modification was the use of an elitist strategy.

Genetic algorithms are stochastic procedures, just as simulated annealing procedures are stochastic. Based only on fitness proportionate selection the population tends to converge to an individual in the initial population (probably one of the fittest). However, this would not be helpful, as the parameter space would not be searched sufficiently. The searching power is provided by the mating and mutation operators. A genetic algorithm is run successfully when a good balance between selection and search has been used. The computational efficiency of the algorithm comes from what can be called *implicit parallelism*. Schemata, which are patterns specified by individual chromosomes, are processed at a much higher rate than the individual strings themselves, because an individual chromosome specifies many schemata. For a population of N chromosomes, an estimate of the number of schemata that are processed in a given generation is N^3 (Goldberg, 1989, pp. 40–41).

The tract variables that were studied here are listed in the first column of Table I. The second column shows the possible variable task-dynamic parameters without the constraints that were applied during recovery. The third column shows the constrained parameters that could be coded into the chromosomes. Usually, not all possible parameters were coded into the chromosomes, and those that were will be made explicit on a case-by-case basis. The constraints were as follows. Constriction location–degree pairs had the same GNFs, and these were equal to the inverse of the durations of the activation intervals (the GNF constraint). The two constriction location–degree pairs were lip aperture, LA, and lip protrusion, LP, and tongue body constriction location, TBCL, and tongue body constriction degree, TBCD. All the activation intervals had to be at least 100-ms duration, and all tract variable motions were critically damped.

The starting activation times and activation durations for both the data-producing task dynamics and the recovered task dynamics were constrained to be in 10-ms increments. Four bits were sufficient to code the starting activation times and activation durations, assuming 10-ms increments. Six bits were used for each of the tract variable targets. Because the resolutions, or discretization increments, of these quantities depend on the allowed ranges of variation, these quantities will be given on a case-by-case basis.

II. THE EXPERIMENTS

There were two groups of experiments performed to study the effects of model mismatch. The first group involved a detailed, qualitative look at a bilabial approximation gesture, including changes in allowed tract variable activation and the amount of acoustic data used in the recovery. Based on the results of these experiments, a more quantitative study was undertaken. This second group of experiments was a systematic study of the effects of mismatch for bilabial and velar approximations.

In the first group of numerical experiments, the weights of the articulators used in the coordinated movement of bilabial approximation, that is, the upper lip, lower lip, and jaw, were altered so that the task dynamics used in recovery was not the same as that used in producing the acoustic data. With these mismatched weights, it would be expected that each articulator trajectory would not be recovered correctly. On the other hand, it might be expected that the task dynamics of the lip aperture tract variable, LA, could be recovered through compensatory activity of the articulators specific to lip aperture (upper lip, lower lip, and jaw). The reason for this expectation is that it can be argued that the acoustic output is sensitive to the task-dynamic specification, because this is a specification of constriction dynamics and the analysis-by-synthesis is based on acoustic matching. If this result attains, then the recovery process would demonstrate articulatory compensation. [See, e.g., Kelso *et al.* (1984) for this concept applied to human speech.] However, it could also happen that task dynamics is not sufficiently constrained so that there is compensation in the tract-variable trajectories (i.e., tract-variable compensation) to produce acoustic outputs similar to those of the data. That is, not only would there be differences between the data producing utterance and the recovered utterance in terms of articulator trajectories, but also in terms of tract-variable trajectories. For instance, in recovery using data produced by a lip approximation, constrictions could be formed at locations other than at the lips, say, at the velum using the tongue body. When this possibility is applied to human speech it is called motor equivalence for acoustic compensation [e.g., Perkell *et al.* (1993); Maeda (1990)].

In the second group of experiments, recoveries of both bilabial and velar approximations were attempted with various articulator weight and/or GNF constraint mismatches. Also, for comparison's sake, an initial study of sensitivity of task dynamics to changes in parameters was undertaken. This was done by synthesizing utterances that differed from designated basic bilabial and velar approximations only in articulator weights and/or GNF. The results of this second group are discussed in terms of articulatory and tract-variable compensation. The robustness of the recovery method was also studied here.

A. First group of experiments

1. Procedure

In the first group of experiments, a bilabial approximation was used to produce the acoustic data consisting of three formant trajectories. The LA was actively reduced from neutral during the first 100 ms, and then it was released to return to neutral position. Lip protrusion was given a target close to its neutral position in the first 100 ms, so that this tract variable did not change much. The weighting of each of the articulators, ULV, LLV, and JA, was set equal to 1 in this data-producing utterance.

During recovery, either the lips (LA and LP), tongue body (TBCL and TBCD), or both were allowed to activate, depending on the recovery type. For either constriction pair, the starting activation time and activation duration were allowed to vary. For the lips, the target for LA also varied, but the LP target was always assumed to be fixed to the correct value. When the tongue body was activated, both the TBCL and TBCD targets varied in the genetic algorithm. The articulator weights were fixed, but sometimes at values did not match those that produced the acoustic data. The resolutions and ranges of the coded target positions are given in Table II.

TABLE II. Target value specifications for the first group of experiments.

Tract variable	Maximum/minimum target value	Number of bits in chromosome	Resolution
LA	23.3/−8.3 mm	6	0.50 mm
TBCL	3.1.6/.51 rad	6	0.042 rad
TBCD	21.5/−6.5 mm	6	0.44 mm

There were three types of recovery attempted with this data. One was where there was the option of activating either, or both, the lips or the tongue body, and the fitness function was based on how well the three formant trajectories fit the data in a mean-square sense [per Eq. (7)]. The second type of recovery was the same, except only the second formant trajectory was used to compute mean-square error. In this case the fitness was calculated using a modified version of Eq. (7), where the formant number index was fixed at 2. A modification to the fitness function using amplitude information was also tried here. Thus the first and second types of recovery differed only in the amount of acoustic information used. In the third type of recovery, only the tongue body task space was allowed to activate. This type was used to test whether tongue body constrictions could be used to compensate for a lip constriction. Within each type of recovery, there were various conditions, where articulator weightings were fixed at different values to test the effect of mismatched weightings on the recovery.

In the first and second type of recovery there was the option of either activating the lips or the tongue body (not exclusive). This was accomplished by adding two bits to each chromosome that indicated whether the lips (LA and LP) or tongue body (TBCL and TBCD) was to be activated (McGowan, 1993). Within these types, there were four recovery conditions: default, where the articulatory weightings of the recovered utterance matched those that produced the data (all articulators weighted at 1), and three conditions with mismatched articulatory weighting for the recovered utterance. In the first mismatch condition the ULV was weighted at 50 and the LLV and JA were weighted at 1, while the second mismatch condition was with the LLV weighted at 50 and the ULV and JA weighted at 1, and the third condition was with the JA weighted at 100 and LLV and ULV weighted at 1. The jaw was an articulator used by both the lip and tongue body task spaces, so the weight of 100 was due to the weighting of 50 from each constriction location–degree pair. Two of the articulators used for the tongue body, the polar coordinates of the tongue body center, CL and CA, were always weighted at 1 in these two types of recovery.

In the third type of recovery, the tongue body constriction degree and location was the only constriction location–degree pair allowed to be activated in the recovery of what had been an articulation produced by the lips. Within this

recovery type, three different conditions were attempted. The first was the default condition with the tongue body vector length, CL, and angle, CA, and the jaw angle, JA, all weighted at 1. The second condition was with CL and CA weighted at 50 and JA at 1. Finally, in the third condition, JA was weighted at 50 and CL and CA were weighted at 1.

For all types and conditions of recovery, the genetic algorithm was run eight times. This allowed for a comparison of eight recovered utterances within each condition. The best of the eight is known as the optimum recovered utterance, and the remaining seven are known as suboptimum recovered utterances. All recoveries were performed with a random initial population of 100 and each was run for 60 generations.

2. Results and discussion

For the first type of recovery and across conditions, none of the 32 recoveries showed activation of the tongue body, and all showed early activation of the lips for a lip aperture reduction target. (Note that tongue body activation can be used to approximate the lips because JA can change with tongue body activation.) In the articulator weight mismatch conditions there was clear evidence of articulatory compensation. The optimum recovered utterances in the three mismatch weighting conditions tracked the data-producing LA trajectory very closely (Fig. 2), despite large differences in articulator trajectories. The mismatch recovery condition with the heavily weighted ULV is taken as an example of articulator differences with the data-producing utterance. Figure 3 shows, that compared to the data-producing ULV trajectory, that there was almost no movement of the upper lip when it was heavily weighted. The compensatory activities of the lower lip, LLV, and jaw, JA, used to attain a reduction in LA, are shown in Figs. 4 and 5.

For the second type of recovery, the acoustic information available for recovery was reduced from the first three formant trajectories to the just the trajectory of the second formant. When the amount of acoustic information was reduced, the number of possible kinds of recovery was increased. In all conditions, default and mismatched articulator weights, there were some of the eight suboptimum recoveries that activated the tongue body and lips, as well as some recoveries that activated the lips alone. In three out of four conditions (default, heavily weighted upper lip, and heavily weighted lower lip), the optimum of eight recoveries were with the lips activated alone to reduce the lip aperture, just as in the first type of recovery. In the case of the heavily weighted jaw, the best recovery involved activating both the tongue body and lips. Also, in the three recoveries where lip activation alone provided the optimum recovered utterance, there was a suboptimum recovery where both the tongue body and lips were activated that also provided a very good fit to the second formant trajectory of the data.

There seemed to be two types of recovery when the tongue body was activated. In the case that ULV was heavily weighted, LA did decrease, but did not track the original data all that well, and the tongue body was activated to form a tight constriction. Visualization on the Haskins articulatory synthesizer graphics revealed that the TBCL was velar. In the case that JA was heavily weighted, the LA actually increased. In fact, the lips were activated to open from 30 to 240 ms in this recovery. Again, the tongue body was activated to make a tight constriction. However, visual inspection revealed that the TBCL was lower in the vocal tract than velar, in what may be described as a uvular or a pharyngeal

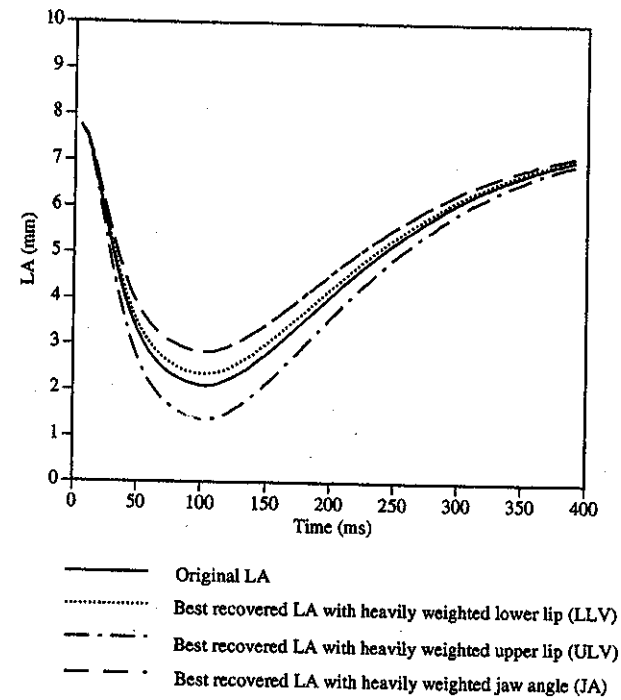

FIG. 2. Optimum lip aperture (LA) trajectory recoveries when both lip and tongue body task spaces were allowed to activate and three formant trajectories were used for acoustic data. Recoveries in various articulator weight conditions are shown.

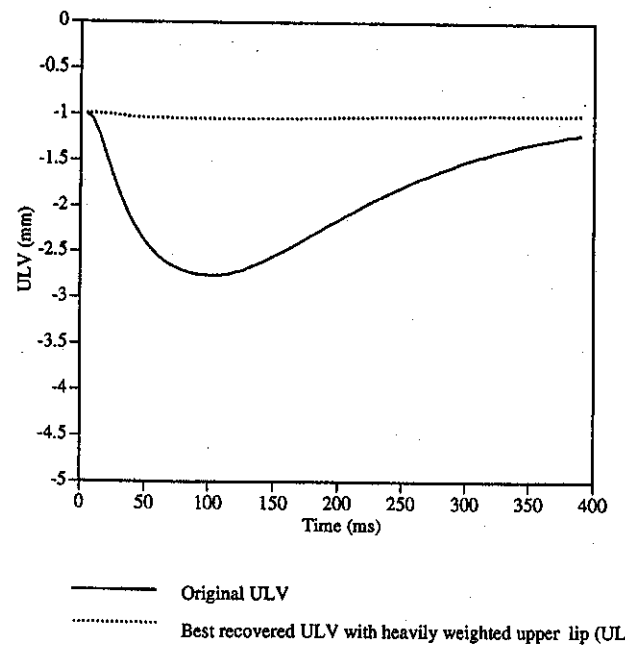

FIG. 3. Upper lip vertical (ULV) trajectories. Shown are the trajectory of the original data and of the optimum recovery with a heavily weighted upper lip when both the lip and tongue body task spaces were allowed to activate and three formant trajectories were used for data.

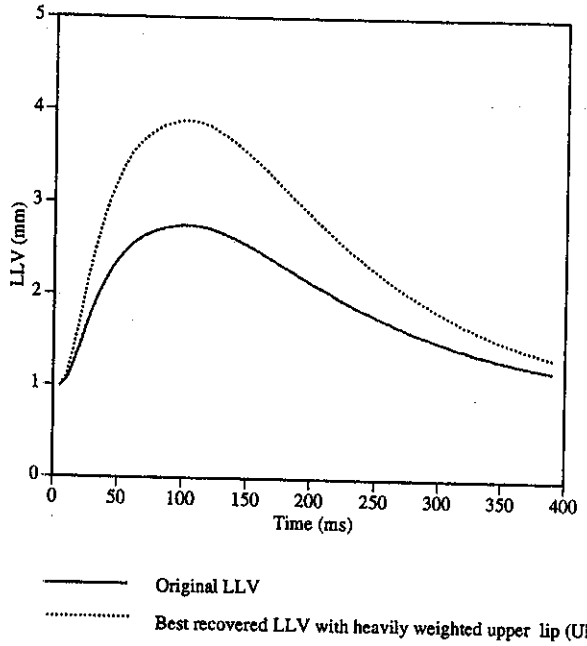

FIG. 4. Lower lip vertical (LLV) trajectories. Shown are the trajectory of the original utterance and of the optimum recovery with a heavily weighted upper lip when both the lip and tongue body task spaces were allowed to activate and three formant trajectories were used for data.

location. A closer look at the formant trajectories produced by these recovered articulations showed that the recovery activating both the lips and the tongue body with a heavy ULV had a lower first formant minimum than did the like recovery with a heavy JA. This is to be expected because the former recovery had closer lips and a more forward tongue body constriction. The second formant trajectories for the two recoveries were similar because, where the closer lips in the heavy ULV recovery tended to produce a lower second formant than in the heavy JA condition, TBCL was more forward for the heavy ULV condition than in the heavy JA condition. This more forward constriction would raise the second formant frequency (see Fant, 1970, p. 121 for a discussion). Neither recovery produced as small a lip aperture as the original data-producing articulation, and both used the tongue body constriction to attain very closely matched second formant trajectories.

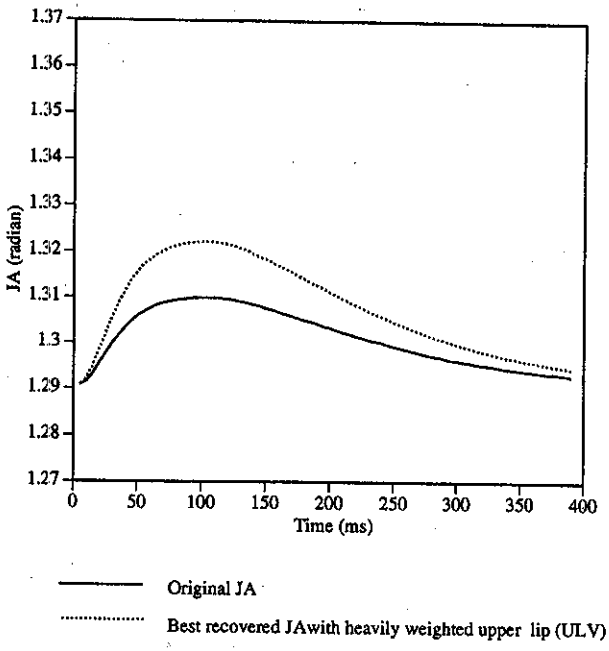

FIG. 5. Jaw angle (JA) trajectories. Shown are the trajectory of the original utterance and of the optimum recovery with a heavily weighted upper lip when both the lip and tongue body task spaces were allowed to activate and three formant trajectories were used for data.

Because tight tongue body constrictions were necessary in the cases when these were used to compensate for a lack of lip closure, there were great amplitude reductions in the speech when this occurred. To test the idea that amplitude information could be used to reduce the number of recoveries involving tongue body activation, the sum of squares of the amplitude differences (in dB) between the speech data and that of the a proposed solution was used in the computation of the fitness. To put the second formant and amplitude on equal footing, the differences of second formant frequency were normalized by the sum of data second formant frequency and the proposed recovered second formant frequency, and the differences in amplitude were normalized by the sum of amplitudes. [Again, the fitness function described in Eq. (7) is modified so that the formant index is fixed at two and a similar sum is taken in amplitude over 10-ms intervals.] After running the second type of recovery under these conditions it was found that only recoveries that activated the lips were obtained, much as for the first type of recovery using the first three formant frequencies as data.

The third type of recovery was where only the tongue body was allowed to activate and the first three formants were used for acoustic data. The recovered utterances had nonconstant lip aperture trajectories, because the jaw moves as an articulator for the tongue body. While these lip aperture trajectories were not as consistent as in the first class of recoveries, there appeared to have been some tract variable compensation with a tongue body constriction: The less approximated the lips, the tighter the tongue body constriction. However, none of the recovered trajectories provided a very good fit, because the recovered third formants moved in opposite direction to those of the data and the recovered second formants did not decrease enough during maximum constriction. Overall, the best recoveries of the third type were not as good as those of the first type in terms of fitness, and these cannot be considered instances of tract-variable compensation.

3. Conclusion

An automatic articulatory recovery procedure has been considered with bilabial approximations under conditions of various mismatched articulator weights between the data-producing task dynamics and the recovering task dynamics. One of the two main results is that the lip aperture was recovered well when there was sufficient acoustic information available and the lips were allowed to activate. The other is that there was no compensation using the tongue body, and the articulators associated with lip aperture compensated for one another in attaining the lip aperture trajectory of the original utterance when there was sufficient acoustic data.

This result provides an instance supporting the work indicating that constriction location and degree are the most acoustically salient aspect of vocal tract shape, particularly when attempting articulatory recovery (e.g., Boë *et al.*, 1992; Papçun *et al.*, 1992). However, when the acoustic information is impoverished, a kind of tract-variable compensation can be seen in the recovery process, with the tongue body forming tight velar, or lower, constrictions, to compensate for less lip aperture decrease. These results in no way inform us on how humans may compensate for either variability in their own articulation or perturbations applied externally. Nor is it certain that tract-variable compensation would not occur in model tests such as these with vocal tract activity other than lip approximation, where the lip protrusion target is assumed known, or with more severely mismatched models. However, these findings encouraged continuation to the second group of experiments to explore more possibilities.

B. Second group of experiments

Because of the interesting qualitative trends noted in the first group of experiments, a more systematic study of bilabial and velar approximations was undertaken. The first part was a sensitivity study, where the effects of perturbing either, or both, the weights of various articulators and/or the GNFs were assessed. In the second part of this group of experiments, a study of the recovery method was undertaken, where either the fixed articulator weights of the recovered utterance mismatched those of the data-producing utterance, the data-producing utterance did not meet the GNF constraint that the recovered utterance was forced to meet, or both. When the data-producing utterance did not meet the GNF constraint it was impossible to recover both the gestural activation interval and the GNF simultaneously.

First, the utterances that were used for comparison in the sensitivity study, and as data-producing and comparison utterances in the recovery study will be described. For all such utterances, movement from a neutral vocal tract into a constriction were activated in the first 100 ms, and all activations were critically damped. All utterances were produced by gestures with one of the two GNFs: one where the GNF met the constraint for recovery that the GNF equal the inverse duration of the activation interval (10 Hz in these cases), and the other where the GNF was 7.2 Hz. For the utterances involving bilabial approximation, the lips were activated with the target for LA of 2.0 mm and LP of 7.5 mm. For velar approximations, the tongue body was activated with the target of the TBCD of 2.0 mm and the TBCL of 1.92 rad, producing a constriction in the velar region (1.57 rad or 90 deg corresponds to the highest point on the palate, and for angles greater than 1.57 rad the constrictions are behind this location). For both the bilabial and velar approximations, the vocal tract was allowed to relax toward neutral position after the initial 100 ms for a total utterance length of 400 ms. The default articulator weightings were as follows, depending on place of articulation. In the bilabial approximations, the default articulator weight was 1 for the LLV, the ULV, and the JA. For the velar approximations, the default articulator weight was 1 for the tongue body CL, tongue body CA, and JA. In the remainder of the discussion, the bilabial and velar utterances with 7.2-Hz GNF and default weighting are known as *basic utterances*. All other utterances in the following sensitivity study are known as *perturbation utterances*.

1. Sensitivity to articulator weight and GNF perturbations: Procedure

A sensitivity study was undertaken to understand the effects of changing the articulator weighting and the GNFs. The comparisons were made between utterances with otherwise identical task-dynamic specifications (place and degree of constriction targets). All comparisons were in root-mean-square difference for relevant tract variables, relevant articulators, and the first three formant frequencies. For bilabial approximations, the comparisons were made between the basic bilabial and perturbed bilabials with a 7.2-Hz GNF, either with the jaw (JA) weight, or with both the lips (LLV and ULV) weights perturbed. The weights of JA or LLV and ULV were perturbed to values 2, 4, 6, 8, 10, 20, 30, 40, or 50. Also, the basic bilabial utterance was compared to the bilabial utterance with 10-Hz GNF, either with default articulator weighting, with JA weight perturbed to 50, or with LLV and ULV weights perturbed to 50. Similarly, for the velar approximations, the comparisons were made between the basic utterance and velar utterances with 7.2-Hz GNF, either with the jaw (JA) weight, or with both the tongue body vector length (CL) and tongue body vector angle (CA) weights perturbed. The weights of JA or CL and CA were perturbed to values 2, 4, 6, 8, 10, 20, 30, 40, or 50. Finally, the basic velar utterance was compared to the velar approximation with 10-Hz GNF, either with default weighting, with JA weight perturbed to 50, or with CL and CA weights perturbed to 50.

2. Sensitivity to articulator weight and GNF perturbations: Results and discussion

The results of these comparisons are given in Tables III and IV for the bilabials and Tables V and VI for the velars. The rms differences for relevant tract variables are zero in all cases of articulator weight perturbation and no GNF perturbation for both bilabials and velars, except when the lips were heavily weighted in the bilabials (Tables III and V). Even in this exceptional case the rms differences in the tract variable LA are very small compared to the rms differences in the articulators LLV and ULV. These observations verify that the task-dynamic model is able to compute the same tract-variable trajectories when the articulator weights are perturbed. For the articulators, there is a steady growth in the differences in the trajectories between the basic utterance and the utterances with heavily weighted articulators as the weights of those articulators grow. (There was one exception in CL for velar approximations, which behaved more erratically.) However, the rate of growth in the differences diminishes with increased articulator weight, indicating that the heavily weighted articulator(s) is (are) nearly completely immobile as the heaviest weight of 50 is attained. The growth in the differences in formant frequencies also diminishes as the weight increases. It is reasonable that the formant frequencies follow this trend of diminished articulator differences, because formant differences were due solely to articu-

TABLE III. The rms differences with the basic bilabial resulting from perturbing articulator weights.

Heavy articulator	Weight	rms differences JA (rad.)	LLV (mm)	ULV (mm)	LP (mm)	LA (mm)	Formant freqs. (Hz)
JA	2	0.0045	0.24	0.24	0.0	0.0	1.5
	4	0.0076	0.40	0.40	0.0	0.0	2.3
	6	0.0088	0.47	0.47	0.0	0.0	2.5
	8	0.0095	0.51	0.51	0.0	0.0	2.8
	10	0.0098	0.53	0.53	0.0	0.0	2.9
	20	0.011	0.57	0.57	0.0	0.0	3.2
	30	0.011	0.59	0.59	0.0	0.0	3.3
	40	0.011	0.60	0.60	0.0	0.0	3.2
	50	0.011	0.60	0.60	0.0	0.0	3.2
LLV & ULV	2	0.0054	0.30	0.30	0.0	0.0054	2.7
	4	0.010	0.57	0.57	0.0	0.0064	4.5
	6	0.013	0.70	0.70	0.0	0.0066	5.1
	8	0.014	0.78	0.78	0.0	0.0069	5.7
	10	0.015	0.83	0.83	0.0	0.0072	5.6
	20	0.017	0.94	0.94	0.0	0.0077	5.9
	30	0.018	0.98	0.98	0.0	0.0077	6.1
	40	0.018	1.01	1.01	0.0	0.0077	6.3
	50	0.019	1.02	1.02	0.0	0.087	6.3

lator differences, as there were essentially no differences in tract-variable trajectories. These tables indicate, not surprisingly, that the output formant frequency trajectories change with articulator trajectory differences with no tract-variable differences. Thus, while acoustic output maybe very sensitive to tract-variable or constriction trajectories, it is not completely determined by these trajectories.

In order that tract-variable trajectories be maintained under perturbation, it is necessary that articulatory compensation to occur. For bilabial approximations trends in rms articulator differences caused by articulatory compensation can be noted (Table III). The jaw vector used for the synthesizer is 113 mm long (Mermelstein, 1973), and the component of lip aperture due to the value of the JA is 113 mm*sin(JA), which is approximately equal to 113 mm*JA for small JA. Throughout Table III it can be seen that rms difference in the component of lip aperture caused by JA is approximately the sum of the rms differences for LLV and ULV. The rms difference in formant frequencies followed fairly closely the trend in rms difference in JA for both heavily weighted JA and heavily weighted LLV and ULV conditions. This is not surprising, because JA affects a much greater portion of the vocal tract compared to the lips.

TABLE IV. The rms differences between the basic bilabial and bilabials with 10- and 7.2-Hz GNFs in different weighting conditions.

Comparison utterance	rms differences JA (rad.)	LLV (mm)	ULV (mm)	LP (mm)	LA (mm)	Formant freqs. (Hz)
10-Hz GNF with default weighting	0.0012	0.12	0.11	0.10	0.36	2.0
10-Hz GNF with heavy jaw, JA weight=50	0.012	0.70	0.70	0.10	0.36	3.6
7.2-Hz GNF with heavy jaw, JA weight=50	0.011	0.60	0.60	0.0	0.0	3.2
10-Hz GNF with heavy tongue body, LLV & ULV weight=50	0.020	1.00	1.00	0.10	0.36	3.6
7.2-Hz GNF with heavy tongue body, LLV & ULV weight=50	0.019	1.02	1.02	0.0	0.0077	6.3

Table IV reports the comparisons of the basic bilabial utterance and the bilabials with 10-Hz GNFs with three different articulator weights perturbations: default weighting, JA weighted at 50, and LLV and ULV weighted at 50. (The results from bilabials with 7.2-Hz GNFs have been included from Table III for ease of comparison.) The rms differences in tract variables, LP and LA, is the same for all articulator weight perturbations. This is consistent with what was observed in Table III, when there were, at most, very small differences in the tract-variable trajectories caused by articulator weight perturbation. This indicates that all such differences must have been caused by the different GNF perturbation from 7.2 to 10 Hz. Again, the trend appears that the greater the difference in JA, the greater the differences in the acoustic output. Comparisons of the rows of Table IV that correspond to the same articulator weight perturbation, but with different GNFs, reveal that the magnitude of articulator and formant frequency differences is unaffected by the perturbations in GNFs, when there are articulators weighted at 50. Thus, at least for the heaviest weight perturbations, and for the GNF perturbation attempted here, the articulator weight perturbations appear to be the dominant factor in determining formant frequency differences.

Velar sensitivity to articulator weight perturbation is shown in Table V. The rms difference in tongue body vector

TABLE V. The rms differences with the basic velar resulting from perturbing articulator weights.

Heavy articulator	Weight	rms differences JA (rad)	CL (mm)	CA (rad)	TBCL (rad)	TBCD (mm)	Formant freqs. (Hz)
JA	2	0.0095	0.23	0.013	0.0	0.0	4.7
	4	0.018	0.23	0.021	0.0	0.0	8.3
	6	0.023	0.23	0.025	0.0	0.0	9.9
	8	0.024	0.23	0.027	0.0	0.0	10.5
	10	0.024	0.23	0.028	0.0	0.0	10.8
	20	0.028	0.23	0.031	0.0	0.0	11.5
	30	0.028	0.23	0.033	0.0	0.0	11.6
	40	0.029	0.23	0.033	0.0	0.0	11.7
	50	0.030	0.0	0.030	0.0	0.0	10.8
CA & CL	2	0.013	0.23	0.0095	0.0	0.0	8.5
	4	0.021	0.23	0.018	0.0	0.0	13.6
	6	0.025	0.23	0.021	0.0	0.0	16.1
	8	0.024	0.0	0.023	0.0	0.0	14.1
	10	0.028	0.23	0.025	0.0	0.0	19.3
	20	0.031	0.23	0.028	0.0	0.0	24.3
	30	0.032	0.23	0.029	0.0	0.0	26.1
	40	0.033	0.23	0.029	0.0	0.0	26.4
	50	0.033	0.23	0.030	0.0	0.0	26.6

length, CL, is mostly a constant, except for a couple of zeros. This indicates some instability in this coordinate, and this situation is currently being remedied by redefining the tongue body vector origin from the current origin, which is the same as that for the jaw vector, that is, at the condyle. The rms differences in JA and CA are comparable. And, in fact, they appear nearly to switch roles when the heavy weight is shifted from JA to CA and CL. In the former case, the CA differences are larger than the JA differences in the same condition, and in the latter case this is reversed. These facts are not surprising, because both angles have the same effect on the position of the tongue body because the vectors originate from the same origin.

TABLE VI. The rms differences between the basic velar and velars with 10- and 7.2-Hz GNFs in different weighting conditions.

Comparison utterance	rms differences JA (rad)	CL (mm)	CA (rad)	TBCL (rad)	TBCD (mm)	Formant freqs. (Hz)
10-Hz GNF with default weighting	0.0032	0.20	0.0032	0.014	0.46	2.9
10-Hz GNF with heavy jaw, JA weight=50	0.026	0.20	0.030	0.014	0.46	11.6
7.2-Hz GNF with heavy jaw, JA weight=50	0.030	0.0	0.030	0.0	0.0	10.8
10-Hz GNF with heavy tongue body, CA & CL weight=50	0.044	0.56	0.027	0.014	0.46	26.6
7.2-Hz GNF with heavy tongue body, CA & CL weight=50	0.033	0.23	0.030	0.0	0.0	26.6

Many of the same trends as were seen for the bilabials hold for the differences between the basic velar utterance and the velar utterances with 10-Hz GNF in various weight conditions (Table VI). The tract variable differences, TBCL and TBCD, were due solely to GNF differences. Also, for the most heavily weighted articulator conditions, the differences in formant frequencies appear to be dominated by differences in articulator weights, and not by differences in GNF. In both the bilabial and velar cases, perturbation to the articulator weights of magnitude 50 appeared to be more destructive to formant frequency matching than did the perturbation of the GNF from 7.2 to 10 Hz. The articulator weight perturbation could be considered to be an articulator perturbation, because these perturbations barely affected the tract-variable trajectories. On the other hand, the GNF perturbation, while certainly changing articulator trajectories, was also a tract-variable perturbation. In summary, the fact that articulator weight perturbations produce formant frequency changes, but not tract-variable changes, makes it possible that tract-variable compensation could occur under articulator weight mismatch model conditions. That is, the optimum acoustic match using a model with articulator weights that do not equal those of the data-producing utterance may be with tract-variable trajectories that are altered to compensate for the altered articulator trajectories. The following study explores this possibility.

3. Recovery with mismatches in articulator weight and/or GNF constraint: Procedure

In these experiments, the recovery algorithm was run enforcing various mismatches in articulator weights between the data-producing and the recovered utterance and/or enforcing mismatches where the data-producing utterance did not meet the GNF constraint that the recovered utterance had to meet. Based on the sensitivity study, the former mismatch

TABLE VII. Target value specifications for the second group of experiments.

Tract variable	Maximum/minimum target value	Number of bits in chromosome	Resolution
LA	1.63/−0.13 cm	6	0.028 cm
LP	1.80/−0.30 cm	6	0.033 cm
TBCL	3.16/0.51 rad	6	0.042 rad
TBCD	1.63/−0.13 cm	6	0.028 cm

may be considered more an articulatory mismatch, and the latter a tract-variable mismatch. In all cases, the recovered utterance had fixed weights corresponding to the default weights used in the sensitivity study, with the exception that one or two of the articulators were weighted at 50. For bilabial approximation recoveries, either JA or LLV and ULV were weighted at 50, and for recoveries of velar approximations, either JA or CA and CL were weighted at 50. A weight of 50 essentially demobilized an articulator so that it played essentially no role in attaining the tract-variable targets. In all cases of recovery the GNF constraint was applied: The GNF had to equal the inverse of the duration of the activation interval.

For both places of articulation, there were three different recovery types, depending on the utterance that produced the acoustic data. Recovery type A was with speech produced by the basic utterance. Recovery type B was with data-producing utterances with a 7.2-Hz GNF and the same articulator weighting as assumed for the recovered utterances, that is, with one or two articulators weighted at 50. Recovery type C was with data produced by utterances with a 10-Hz GNF and default weighting, thus meeting the GNF constraint imposed for recovered utterances. The recovered utterances were always compared to the utterance that produced the data, as well as to the corresponding basic utterance.

The recoveries were run 32 times for each recovery type and for each heavy articulator. The ranges and resolutions of the relevant tract-variable target specifications in the genetic algorithm are given in Table VII. Note that for these bilabial recoveries that the LP target was recovered, instead of being assumed known as in the first group of experiments. For the data created by bilabial approximation, the recovery was done assuming that only the lips, LA and LP, could activate, and for data produced by velar approximations, only the tongue body, TBCL and TBCD, could activate in a recovered utterance. Otherwise, the genetic algorithm parameters were set as in the first group of experiments.

4. Recovery with mismatches in articulator weight and/or GNF constraint: Results and discussion

The best of 32 recovered utterances in each condition and place of articulation was taken for comparison with its data-producing utterance. As before, the fittest of the 32 recovered utterances is called the optimum recovered utterance, and the remaining 31 are called suboptimum recovered utterances. Comparisons were done using root-mean-square differences in formant frequencies, articulatory trajectories, and tract-variable trajectories, just as in the sensitivity study.

TABLE VIII. The rms differences between the basic bilabial and the optimum recovered bilabial utterance in three conditions with different articulator weight mismatches.

		rms differences					
Heavy articulator	Recovery type	JA (rad)	ULV (mm)	LLV (mm)	LP (mm)	LA (mm)	Formant freqs. (Hz)
JA	A	0.015	1.03	1.03	0.43	0.18	0.96
	B	0.018	1.03	1.03	0.85	0.52	6.15
	C	0.011	0.86	0.86	0.44	0.32	3.34
ULV & LLV	A	0.011	0.76	0.76	1.77	1.10	2.68
	B	0.011	0.16	0.16	0.035	0.079	3.22
	C	0.031	1.06	1.06	1.77	1.01	1.59

The results from recoveries can be compared to the sensitivity results where the corresponding articulators were weighted at 50 (Tables IV and VI).

In type A recoveries there was a mismatch in articulator weight with the data-producing utterance and the data-producing utterance did not meet the GNF constraint. The data-producing utterances were the basic utterances in this case. Despite the severity of the mismatching, in three out of four cases the rms difference in formant frequency was not the largest of the three recovery conditions (Tables X and XI). Also, as will be discussed below, the formant frequency differences in these optimum recoveries with heavily weighted articulator(s) and constrained GNFs were always less than the formant differences caused by perturbing the same articulator weight(s) of the basic utterance to 50 and perturbing its GNF to 10. Some kind of compensation must have been used to obtain these results. Because the same tract-variable trajectories as those of the data-producing basic utterance were not attained, some of this compensation might be attributed to tract-variable compensation.

A closer look at the bilabial type A recoveries is revealing. For bilabials with the jaw angle JA weighted at 50, comparing the first row of Table VIII with the second and third rows of Table IV shows rms articulator differences (JA, LLV, ULV) of the optimum recovered utterance with the basic utterance a little greater than the differences between perturbed utterances and the basic utterance, but of comparable magnitude. Also, the utterance perturbed from the basic utterance with a JA weight of 50 and a GNF of 10 (row 2 of Table IV) has comparable rms differences in LA and LP with those of the optimum recovered utterance, which also had a 10-Hz GNF, allowing that LA and LP can offset one another. However, the optimum recovered utterance was closer to the basic utterance in terms of formant frequency than either of the perturbed utterances, 10 or 7.2 Hz, with the jaw angle weight perturbed to 50. The story is generally similar in comparing the recovered and perturbed utterances with the lips weighted at 50 (the fourth row of Table VIII can be compared with the fourth and fifth rows of Table IV), except that the optimum recovered utterance, which had a 10-Hz GNF, had LP and LA trajectories with relatively large rms differences with the basic trajectories compared to the perturbed utterance with 10-Hz GNF. Again, the optimum recovered utterance had a better frequency match with the ba-

TABLE IX. The rms differences between the basic velar and the optimum recovered utterance in three conditions with different articulator weight mismatches.

Heavy articulator	Recovery type	rms differences JA (rad)	CL (mm)	CA (rad)	TBCL (rad)	TBCD (mm)	Formant freqs. (Hz)
JA	A	0.030	1.80	0.076	0.11	0.21	7.7
	B	0.030	0.15	0.028	0.0058	0.096	10.8
	C	0.030	2.21	0.0029	1.01	0.28	8.3
CA & CL	A	0.023	2.45	0.030	0.046	1.06	5.6
	B	0.030	0.26	0.030	0.0054	0.14	22.8
	C	0.030	2.21	0.0029	1.01	0.28	8.3

TABLE X. The rms differences between the data-producing bilabial and the optimum recovered bilabial utterance in three conditions with different articulator weight mismatches.

Heavy articulator	Recovery type	rms differences JA (rad)	ULV (mm)	LLV (mm)	LP (mm)	LA (mm)	Formant freqs. (Hz)
JA	A	0.015	1.03	1.03	0.43	0.18	0.96
	B	0.0	1.39	1.39	1.46	0.94	0.48
	C	0.012	0.78	0.78	0.72	0.29	3.0
ULV & LLV	A	0.011	0.76	0.76	1.77	1.10	2.7
	B	0.0021	0.0040	0.0059	0.062	0.17	1.8
	C	0.0023	1.09	1.09	3.19	1.94	1.4

sic trajectory than the two perturbed utterances.

Despite the mismatches, it was possible to have the optimum recovered utterances agree closely with the basic utterance in formant frequencies because task-dynamic specifications of the optimum recovered utterances, such as the target positions and the activation interval, could be used to compensate for the differences in articulator weight and GNF with the basic utterance. The basic bilabial utterance had a GNF of 7.2 Hz, lip activation from 0 to 100 ms, an LA target of 2.0 mm, and an LP target of 7.5 mm. Both optimum recovered utterances (heavy JA and heavy LLV and ULV) had activation intervals from 10 to 110 ms, and, therefore, a 10-Hz GNF. The optimum recovery for the heavy jaw mismatch produced an LA target of 1.8 mm and an LP target of 6.33 mm, and the optimum recovery for the heavy lip mismatch produced an LA target of 5.41 mm and an LP target of 12.67 mm. Thus these recoveries appear to be a case where recovered task-dynamic parameters necessarily differed with those of the data-producing utterance, yet matching the acoustic data very closely (note the LA and LP columns in Table VIII). The bilabial approximation may be particularly prone to tract-variable compensation, because, for example, any lack in lip closure could be compensated for with extra lip protrusion. In other words, not only can the dynamics of the tract variables compensate for mismatches in task-dynamic specification of the data-producing utterance, but the tract variables can compensate for one another, even without mismatch. Note that greater or lesser LA target meant greater or lesser LP target for the two optimum recovered utterances.

The type A optimum velar recoveries also show better matches in formant frequencies with the basic utterance than the corresponding perturbed utterances did (compare the first row of Table IX with the second row and third rows of Table VI, and the fourth row of Table IX with the fourth and fifth row of Table VI). This is in spite of the fact that the optimal recovered utterances did not do any better matching the basic velar's tract variable, TBCL and TBCD, trajectories than did the perturbed utterances. The two optimum recovered velar utterances had GNFs of 8.33 Hz, as compared to 7.2 Hz for the basic utterance. Thus the optimum activation interval was lengthened from 100 to 120 ms in both the heavy jaw and the heavy tongue body mismatch conditions. For the basic utterances the TBCL target was 1.92 rad and the TBCD target was 2.00 mm, while the optimum recovered utterance with a jaw weight mismatch had a TBCL target of 2.09 rad and a TBCD target of 2.08 mm, and the optimum recovered utterance with the tongue body heavily weighted had a TBCL target of 1.81 rad and a TBCD target of 5.14 mm. Again, it appears that task-dynamic specification differences were used to produce tract-variable trajectories differing from the data-producing utterance, yet achieving good formant trajectory matching.

The other two recovery types were from data produced by utterances that were not as severely mismatched in terms of articulator weight or GNF constraint. For recovery type B, the utterance that produced the data did not obey the GNF constraint with activation intervals 100 ms long and a 7.2-Hz GNF, but the articulator weights were identical to those of the recovered utterances. In fact, given the results of type A recoveries, where the mismatches were more severe, it would be expected that the optimum recovered utterances in this condition would have little trouble in matching the formant trajectories of the data-producing utterance. And, in every case, the type B recovery did provide a tighter acoustic fit than did the type A recovery (Tables X and XI).

The optimum recoveries of bilabials again showed the LP and LA targets compensating for one another. In the heavy jaw condition, the optimum recovered the rms differences for both LA and LP were greater in the type B recovery than in the type A recovery, despite the fact that the rms formant differences were smaller in the type B recovery (Table X). This suggests that the two lip tract variables are trading with one another to attain good acoustic fit. In contrast to type A, the velar TBCL and TBCD trajectories were recovered very well in type B recoveries in both the heavy jaw and heavy tongue body conditions (Table XI). And, the reduction in formant frequency difference might be attributed to a better tracking of tract variables in velar type B recoveries than for velar type A recoveries. However, the type B recoveries also showed less rms difference in articulator trajectories, so the close acoustic fit could also be attributed to the close articulator fit.

For type C recoveries, as with type B recoveries, the optimum recovered utterances should have had little trouble in tracking the data's formant frequencies. This is because the data-producing utterance mismatched the recovered utterances in articulator weight, and not GNF constraint, and thus

TABLE XI. The rms differences between the data-producing utterance and the optimum recovered utterance in three conditions with different articulator weight mismatches.

Heavy articulator	Recovery type	rms differences JA (rad)	CL (mm)	CA (rad)	TBCL (rad)	TBCD (mm)	Formant freqs. (Hz)
JA	A	0.030	1.80	0.076	0.11	0.21	7.7
	B	0.00035	0.28	0.0074	0.010	0.17	2.2
	C	0.028	1.91	0.0010	0.17	0.35	9.0
CA & CL	A	0.023	2.45	0.030	0.046	1.06	5.6
	B	0.0020	0.23	0.0	0.0077	0.24	3.2
	C	0.028	1.91	0.0010	0.17	0.35	9.0

there was less mismatch than in the type A recovery. Although type C recoveries had the worst formant frequency match in three out of four cases, formant frequencies differences were all within the magnitude of the other recovery types (Tables X and XI). Also, considering the velar recoveries, the type C recoveries had larger differences in tract variables, TBCL and TBCD, than did the type B, which corresponds to the differences in the formant frequencies in the two mismatch weight conditions (Table XI). This was also true of the type C compared to the type A recoveries, except that the rms difference in TBCD was greater for the type A than the type C recovery for the heavy tongue body condition (Table XI). It is interesting to note that the algorithm had a harder time overcoming mismatches in articulator weight (type C) than it did GNF constraint (type B) to match formant frequencies, despite the fact that the former was not a tract-variable mismatch. This parallels the observation that the articulator weight perturbations in the sensitivity study was more destructive to formant frequency matching than the GNF perturbation.

Finally, the distribution of suboptimal solutions and the shape of the fitness function in the neighborhood of the optimum solution are considered. A way to visualize this is to plot the *task-dynamic distance* of each of the suboptimum recovered utterances from the optimum recovered utterance as a function of their fitness. The task-dynamic distance between two utterances is defined to be the number of increments between the task-dynamic parameters in the two utterances, where an increment for a given parameter is the resolution of that parameter defined in the genetic algorithm coding. For utterances using a single task-dynamic constriction location–degree pair (place of articulation), with a starting activation time, duration of activation, and two target positions to be coded, the task-dynamic distance between two utterances would be the number of 10-ms increments difference in their starting activation times plus the number of 10-ms increments difference in their activation durations, and the number of increments difference in both the targets. The increments for the targets are determined by the resolutions given by the chromosomal coding, and these are specified in Table VII. Thus for a bilabial approximation recovery the distance between the optimum recovery and one of the suboptimum recoveries would be the number of 10-ms increment differences in the start of lip activation, plus the number of 10-ms increments in the duration of the activation, plus the number of increments difference in the LA target, plus the number of increments difference in the LP target. The resulting distances were plotted against the utterance fitness to assess the regularity of the fitness landscape near the optimum recovered utterances. Further, to see whether getting close to the optimum recovery was a common or rare event, histograms of the number of suboptimum recoveries were also plotted. The range of fitness from the minimum suboptimum fitness to the optimum fitness was divided into 20 equal-fitness intervals to generate the histograms (see Figs. 6 and 7 for examples). Two measures were derived to quantify the shapes of these plots. The eveness measure for the histograms, the EMH, is the square root of the sum of squares of the differences of the height of the bars from 1.6, which would be the height of each bar of the 32 suboptimal or optimal solutions if they were evenly distributed between the 20 divisions. The measure of roughness of the fitness landscapes, the RFL, is the number of extrema multiplied by the square root of the sum of squares of differences in task-dynamic distances between suboptimal or optimal solutions with neighboring fitnesses (squares of the vertical distances between neighboring points in the task-dynamic distance versus fitness plots), divided by the task-dynamic difference between the suboptimal solution with the least fitness and the optimal solution. The division normalized for the absolute

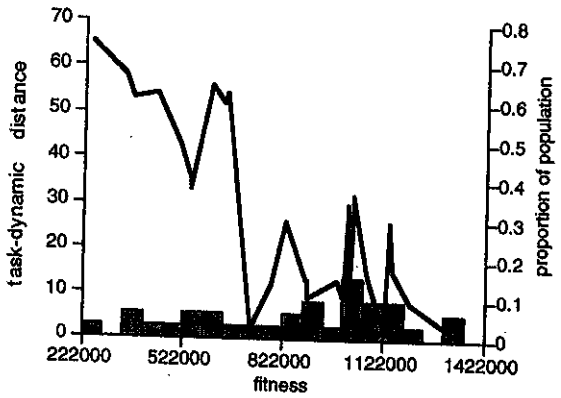

FIG. 6. Task-dynamic distance and proportion of population versus fitness for type B bilabial recovery with a heavily weighted JA.

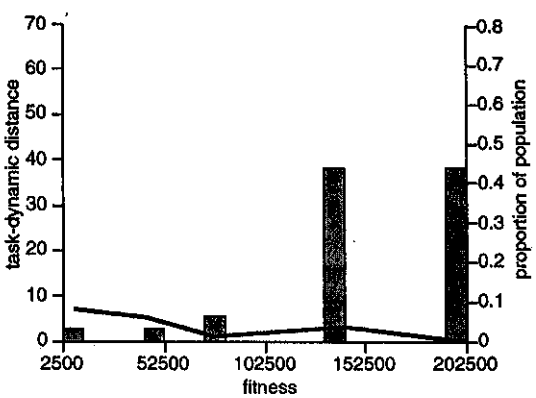

FIG. 7. Task-dynamic distance and proportion of population versus fitness for type B velar recovery with a heavily weighted JA.

descent of the task-dynamic versus fitness curve, so that this did not affect the calculation of roughness.

On the whole, the bilabial suboptimum recoveries were more evenly distributed than were the velar suboptimum recoveries, over all recovery types and conditions. The mean EMH for bilabial recovery was 8.1, with a maximum of 13.9 and a minimum of 4.4, and the mean EMH for velar recovery was 17.8 with a maximum of 23.5 and a minimum of 9.0. The differences in the eveness of the histograms were more pronounced when the comparison was restricted to the heavy jaw condition. The mean EMH for bilabial recoveries across the three recovery types was 8.5 with a maximum of 13.9 and a minimum of 4.6, and for the velar recoveries, the mean EMH was 21.2 with a maximum of 23.5 and a minimum of 17.5. The roughness of the task-dynamic distance versus fitness curves was also greater for bilabial recoveries than velar recoveries in the heavy jaw condition. Across the recovery types and in the heavy jaw condition, the mean RFL for bilabials was 10.8 with a maximum of 20.3 and a minimum of 5.2, whereas for the velars the mean RFL was 2.4, with a maximum of 4.3 and a minimum of 0.7. The differences in EMH (eveness) and RFL (roughness) between bilabial and velar recoveries in the heavy jaw condition are indicative of the compensation that LA and LP can provide for one another during bilabial recovery. Also, there was a tendency of distributing the suboptimum solutions evenly from minimum fitness to maximum fitness for recoveries done from data produced by bilabial utterances that did not meet the GNF constraint (type A and B recoveries) and for velars that did not meet the GNF constraint with a heavily weighted tongue body. (The mean EMH across this set was 7.3, with a maximum of 11.3 and a minimum of 4.4. The mean EMH over the remaining histograms was 18.5, with a maximum of 23.4 and a minimum of 11.1.) For velar recoveries with mismatches in GNF constraint (type A and B recoveries), the suboptimum recoveries with the heavily weighted jaw were less distributed away from the optimum than in the case that the tongue body was heavily weighted (EMHs of 17.6 and 22.5 vs 9.0 and 11.3, respectively). When there is a low number of degrees of freedom, as when the tongue body is heavily weighted and only the jaw can move, mismatches in GNF constraint mean that the genetic algorithm is more likely to get trapped into a local fitness maximum, especially one that is close to the optimum recovery in terms of task-dynamic distance, but not fitness. In other words, once the population of chromosomes starts to converge to the suboptimum solution the diversity in the population is lost, and it is difficult to flip the few bits that are necessary to attain a much better fitness. Thus it is difficult to get to optimum when the suboptimum solutions are close in task-dynamic specification and but not in fitness. The even distributions in the case of bilabials seem, rather, to be caused by the trading possible between the tract variables.

III. SUMMARY

In the first group of experiments, where recovery of lip approximation was performed and where only the lip aperture parameters were treated as unknown, the articulator weights were mismatched with those of the data-producing utterance. Articulatory compensation was evident when sufficient acoustic data was available, that is, when either there were three formant trajectories or the second formant trajectory and amplitude information. In these cases, the optimum recovered articulator trajectories were adjusted according to the discrepancy caused by the misweighted articulator so that the recovered lip aperture trajectory was very close to that of the data-producing utterances. This was no longer the case when only the second formant trajectory was used for acoustic data. In this condition, sometimes the tongue body would be activated in a recovered utterance rather than the lips.

In the second group of experiments both lip and velar approximations were considered. A sensitivity study showed that there could be formant frequency trajectory changes resulting from articulator trajectory changes, even when there were no tract-variable trajectory changes. Not surprisingly, the correspondence between tract-variable (or constriction) dynamics and acoustic output is not exact.

There were three types of recovery attempted in the second group of experiments. Despite the various mismatches in either, or both, the articulator weights and GNF constraints, the formant frequency matches between the optimum recovered utterances and the data-producing utterances were close, particularly considering the results of perturbations corresponding to the mismatches in the recovery process. The tract-variable trajectories of the data-producing utterances were not recovered exactly, and there was evidence that the tract variables, and not articulatory trajectories alone, were used in the compensation to attain a good acoustic match. In the case of bilabial approximation recovery, there was no correspondence between the match in tract variables and the match in formant frequencies in the optimum recovered utterances. This could be caused, partly, by the trading between lip aperture and lip protrusion. On the other hand, for velar approximations, there was some correspondence between the closeness of the tract variables and formant trajectories. Thus, while tract-variable trajectories could be manipulated during recovery to attain good formant matching, various types of recovery for the velar place of articulation showed a correspondence between tract-variable and formant trajectory differences. However, as the sensitivity study and the bilabial recoveries show, there is not, in general, a correspondence between tract-variable trajectory differences and the acoustic differences measured in terms of the first three formant frequencies.

In none of the recoveries in the second group of experiments, particularly in the type C recoveries, where there was only articulator weight mismatch, did the optimum recovered utterance recover the original tract-variable trajectories. (The sensitivity study showed that it was possible to obtain almost exactly the same tract-variable trajectory, even when articulators weights are perturbed.) The reason for this difference is that the recovery was performed in an analysis-by-synthesis to match formants, while task dynamics performs a least-squares projection onto the articulator trajectories to obtain the correct tract-variable trajectories. This is one effect of the pseudoinverse of the Jacobian in Eq. (5). It is known from the sensitivity study that the same tract-variable trajectories do not lead to the same three formant trajectories,

but that between recovery types a correspondence between the two specifications is preserved. Also, when the recovery is sufficiently constrained, as in the first group of experiments, there is obvious articulatory compensation in recovery to attain both similar tract variable and formant trajectories. One of the goals of the future research in this method of articulatory recovery is to find the sufficient acoustic information to preserve the correspondence between tract-variable difference and acoustic difference. A more ambitious goal is to make the task-dynamic distance versus fitness landscapes monotonic functions with good acoustic information. It must be borne in mind, however, that the results will be dependent on the particular articulatory synthesizer that is chosen for the analysis-by-synthesis.

ACKNOWLEDGMENTS

This work was supported by the NIH through Grant No. NIH DC-01247 to Haskins Laboratories. The comments of the anonymous reviewers were very helpful.

Boë, L.-J., Perrier, P., and Bailly, G. (**1992**). "The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion," J. Phon. **20**, 27–38.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Forrest, S. (**1993**). "Genetic algorithms: Principles of natural selection applied to computation," Science **261**, 872–878.

Goldberg (**1989**). *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA).

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., and Fowler, C. A. (**1984**). "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures," J. Exp. Psychol.: Hum. Percept. Perform. **10**, 812–832.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, The Netherlands), pp. 131–149.

McGowan, R. S. (**1993**). "Implementing a genetic algorithm to recover task-dynamic parameters of an articulatory speech synthesizer," Haskins Laboratories Status Report on Speech Research No. SR-113, pp. 95–106.

McGowan, R. S. (**1994**). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," Speech Commun. **14**, 19–48.

McGowan, R. S. (**1995**). "Recovering task dynamics from formant frequency trajectories: Results using computer babbling to form an indexed data base," in *Producing Speech: Contemporary Issues*, edited by F. Bell-Berti and L. J. Raphael (AIP, Woodbury, NY), pp. 489–504.

Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. **53**, 1070–1082.

Papçun, G., Hochberg, J., Thomas, T. R., Larouche, F., Zacks, J., and Levy, S. (**1992**). "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," J. Acoust. Soc. Am. **92**, 688–700.

Perkell, J. S., Mattheis, M. L., Svirsky, M. A., and Jordan, M. I. (**1993**). "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot motor equivalence study," J. Acoust. Soc. Am. **93**, 2949–2961.

Rubin, P., Baer, T., and Mermelstein, P. (**1981**). "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. **70**, 1109–1121.

Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamic approach to gestural patterning in speech production," Ecol. Psychol. **14**, 333–82.