Recovering task dynamics from speech acoustics: Numerical results and the application of the method in speech technology

R.S. McGowan

Haskins Laboratories
270 Crown Street
New Haven, CT 06511

Abstract
The problem of recovering speech articulation from the speech acoustic signal, the speech inverse problem, has seen much progress in the last forty years. Much of this progress has resulted from increased knowledge of speech motor control, which may allow the mapping from acoustics to articulation to be constrained sufficiently to be unique and robust. The task-dynamic model of speech production, developed at Haskins Laboratories by Saltzman and his colleagues, incorporates knowledge of speech motor control into a computational model. Task-dynamics is used in the method proposed here as a means to constrain the inverse mapping.

A brief description of the proposed method is given, including task-dynamics and the genetic algorithm that is used in the optimization. Then some results from computer experiments are given. The method is analysis-by-synthesis, where the speech of a proposed articulation is compared to the speech data. The proposed articulations are specified by task-dynamic parameters as coded into chromosomes strings. A genetic algorithm is applied to a population of these strings, so that the speech of the proposed articulations approaches that of the data.

Articulatory recovery, or at least articulatory constraint, has been proposed as a way to perform bit-rate reduction and as part of the way to do automatic speech recognition. Future uses of this method in automatic speech recognition are proposed.

Keywords: inverse problem, articulatory recovery

1. Introduction

Research into methods for recovering vocal tract articulation from speech acoustics has largely been driven by technological applications. The problems of low bit-rate coding and automatic speech recognition are two examples (see e.g. [1,2]). Because changes in articualtory positions occur at a much lower rate than changes in the acoustic pressure wave, the transmission of articulatory information could occur at a much lower rate than the transmission of the wave itself. The utility of vocal tract recovery and constraint in automatic speech recognition will be considered at the end of this paper.

In proposed solutions to the speech inverse problem, there has been a trend of including more articulatory constraints to reduce the number of degrees of freedom in the articulatory space (see e.g., [2,3]). The reason for this trend is to make the mapping from acoustic data to the articulatory domain less ambiguous and less sensitive to noise

in the acoustic data. In the recent history of the problem, the articulatory domain (the range space of the inverse mapping) has progressed from the area function of a tube to computational models of articulation. The schematic midsaggital outline of the articualtory model used in the Haskins articualtory synthesizer, ASY, is shown in Figure 1 [4,5]. Further constraints on the manner that the articulators move have been applied in solutions to the inverse problem. The movements of articulators have been constrained to move smoothly [3] or to move according to some parametric representation [2], so that sequential acoustic data are assumed to specify articulatory sequences that do not change too abruptly.

2. Task-dynamics

The trend of increasing articulatory constraint has been continued in the articulatory recovery method to be described here [6]. Not only is a computational model of articulation used to constrain the articulatory space here, but a model for coordination of the articulators is used: Task-dynamics [7]. Task dynamics models the formation of constrictions with vocal tract articulators. For instance, the tongue tip forms constrictions behind the teeth for /t/, /d/, and /s/, while the constriction for /b/ and /p/ is at the lips. Note that the degrees of constriction are different for the /d/ and the /s/, despite very similar constriction locations. Thus, task dynamics allows the specification of both where and to what degree a constriction is made. Even vowels are described using constriction location and degree, although the degree of constriction for the vowels are less than those for the obstruent consonants. The locations of constrictions and degrees of constriction are specified by *tract variables*. A partial list of tract variables is: tongue body constriction location and degree (TBCL and TBCD), tongue tip constriction and degree (TTCL and TTCD), and lip protrusion (location) and aperture (degree) (LP and LA) (Figure 2).

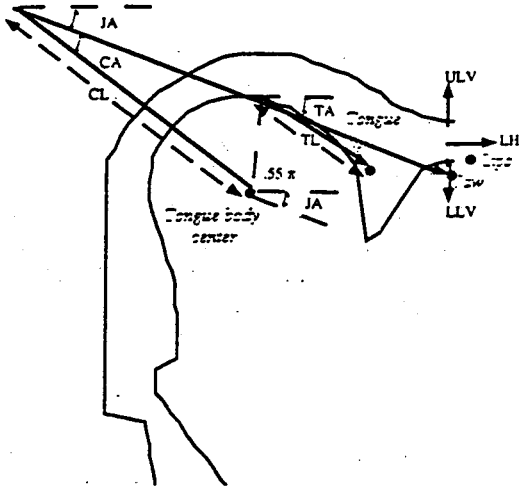

Figure 1. The ASY vocal tract.

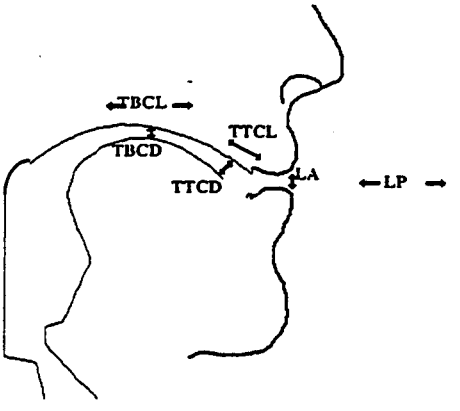

Figure 2. The tract variables

The task-dynamic model assumes that the each tract variables possesses a linear second-order dynamics, so that when a tract variable is activated it has a damping ratio, natural frequency, and a target position that need to be specified.

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

where $\mathbf{z}$ is the vector of tract variables, $\mathbf{M}$ is the diagonal mass matrix, assumed to be identity, $\mathbf{B}$ is the diagonal damping matrix, $\mathbf{K}$ is the diagonal stiffness matrix, and $\mathbf{z}_0$ is the vector of target positions. For example, to form a bilabial closure for a /b/ one might specify a critically damped LA with a natural frequency of 10 Hz and a zero or negative target. The interval over which this specification is active, the *activation interval,* must also be given. For the case of the bilabial closure that might be from zero to 90 ms. Such a closure has been illustrated in Figure 3. Following the trajectory for LA, it can be seen that the lips close, i.e. LA < 0. Figure 3 is a graphical representation of a *gestural score*. The heights of the shaded boxes correspond to target positions, and the horizontal dimensions of the boxes correspond to activation intervals. The trajectories of the tract variables resulting from the task-dynamic specification are shown in each of the panels.

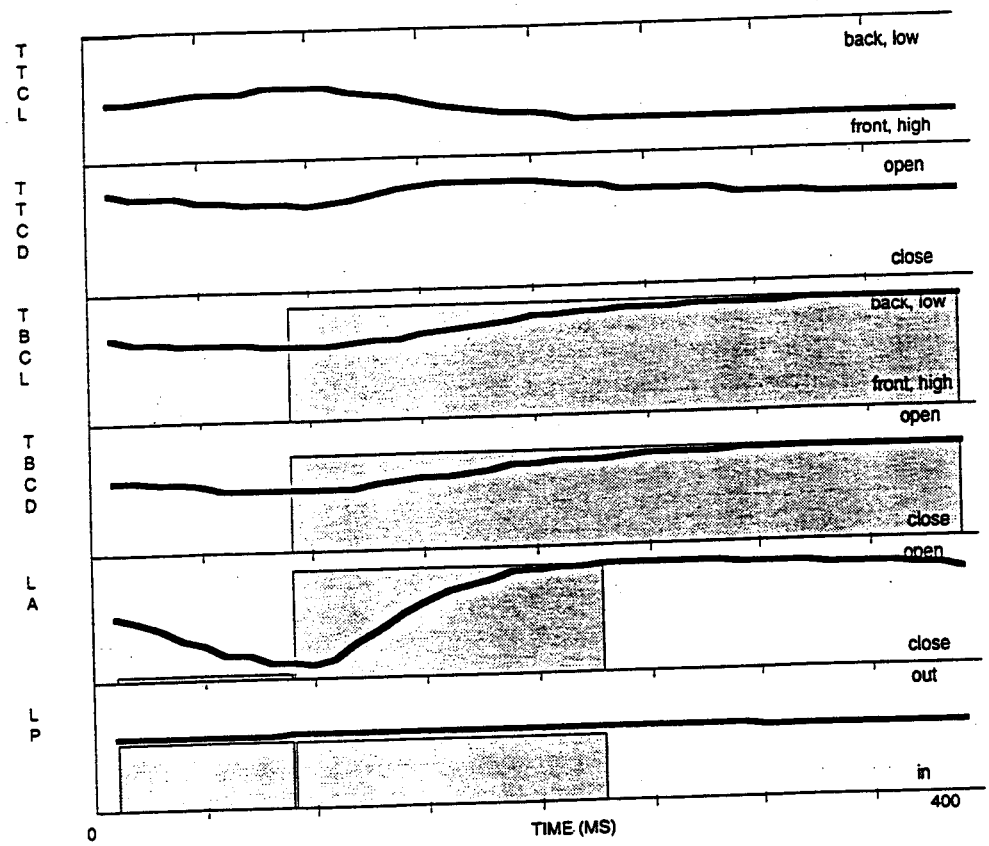

Figure 3. Tract variables trajectories and activations for /əbæ/. The activation times are shown by the length of the shaded boxes and the targets by the heights. The maximum possible value for TBCL is 3.49 radians. The maximum possible value for TTCL is 1.22 radians. The minimum possible value for both TBCL and TTCL is -0.18 radians. All other tract variables, TTCD, TBCD, LA, and LP have -1.0 cm for a minimum possible value and 2.5 cm for a maximum possible value.

Note that frequencies and dampings are not shown in this representation. After the lip closure in Figure 3, the tongue body is activated so that it forms a low (TBCL) and relatively front vowel (TBCD). Because some tract variables depend on common articulators some tract variables move despite the fact that they are not activated. Thus, despite the fact that TTCL and TTCD are not activated, they move because they use the tongue body, as do the activated TBCL and TBCD.

Task dynamics provides an abstract, mathematical description of how constrictions are formed and broken during speech. However, for this description to be of any use for our purposes, it must be instantiated into an articulatory synthesizer, such as the one shown in Figure 1. That is, the task dynamics must be mapped to a dynamics of the articulators of the synthesizer. This is accomplished by specifying a geometric mapping from the articulators to the tract variables.

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

where ϕ is the vector of articulator coordinates. For instance, the tract variable LA in Figure 2 is specified by the articulators jaw angle (JA), upper lip vertical position (ULV), and lower lip vertical position (LLV) in Figure 1. Substituting equation (2) into equation (1), and after some manipulation, a dynamical description in terms of articulatory variables is obtained.

$$\ddot{\phi} = \mathbf{J}^{*}\left\{ \left(\mathbf{M}^{-1}\left[-\mathbf{B}\mathbf{J}\dot{\phi} - \mathbf{K}\Delta \mathbf{z}(\phi)\right]\right) - \dot{\mathbf{J}}\dot{\phi}\right\} \tag{3}$$

where $\Delta\mathbf{z} = \mathbf{z} - \mathbf{z}_0$, $\mathbf{J}$ is the Jacobian of the mapping in equation (2), and the star denotes a weighted pseudoinverse. Note that the set of dynamic equations (3) in the articulator space is a set of nonlinear, coupled differential equations. Assuming that the rows of $\mathbf{J}$ are linearly independent:

$$\mathbf{J}^{*} = \mathbf{W}^{-1}\mathbf{J}^{T}(\mathbf{J}\mathbf{W}^{-1}\mathbf{J}^{T})^{-1} \tag{4}$$

where $\mathbf{W}$ is the weighting matrix and the superscript T denotes transpose. $\mathbf{W}$ 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 is the same for a given articulator no matter which tract variable is involved. Note that the larger the weight, w_{jj}, the smaller the weighted derivative of any tract variable with respect to the j^{th} articulator. The larger the weight of the j^{th} articulator the less likely it is to move to attain tract variable goals.

In sum, not only have we constrained the articulator space to have human-like articulators that move continuously in time, but the dynamics of the articulation has been parameterized. In the recovery method proposed here, it is the set of task-dynamic parameters: damping, natural frequency, target, and activation intervals that are to be recovered from the speech acoustics.

There is an extra bonus in mapping from acoustics to task dynamics because the acoustics are more sensitive to constriction degree and location than to any other part of

the area function. This follows by a consideration of the Webster horn equation, where there is a logarithm of area dependence for the coefficient of the spatial derivative term, and by the experimental results of others [8,9].

To reduce the number of degrees of freedom in the work described here, there were constraints imposed on the task-dynamic parameters. First, all second-order dynamics were assumed critically damped. Further, the natural frequency of any activation was assumed to be the inverse of the duration of the activation interval. Also, constriction location and degree pairs: LA and LP, TBCL and TBCD, and TTCL and TTCD were constrained to have the same activation intervals.

3. ASYINV- a program for task dynamic recovery

A computer program, ASYINV, has been written to test the possibility of recovering the task dynamics of a talker from his or her speech acoustics. The data input are the frequencies of the first three resonances (formant frequencies) of the vocal tract. Because task-dynamic parameters are to be recovered, and the effects of these parameters are over finite durations, the data is also given over finite durations. Each of the three formant frequencies are sampled at 100 Hz for as long as the utterances lasts. (In the initial testing to be reported here this is over the duration of a vowel-consonant-vowel sequence, or about 400 ms.)

The proposed solutions were provided by an optimization procedure that falls under the general heading of analysis-by-synthesis or hypothesis testing. A genetic algorithm, which is more generally an adaptive procedure, was used for optimization [10]. In this procedure, task-dynamic parameters were coded into binary strings called chromosomes. The choice of an initial population of chromosomes is random, and each string, representing a model solution, is assigned a fitness based on the fitness function:

$$fitness = \left(\sum_{i=1}^{3} \sum_{j=1}^{N} \left(f_{ij}^{model} - f_{ij}^{data} \right)^2 \right)^{-1} \qquad (5)$$

where the i subscript denotes formant number and j denotes time step. The population of chromosomes is run through a series of generations of fitness weighted selection, mating, and mutation. The population of chromosomes becomes more homogeneous, while increasing the average fitness of the population. At the end of a prespecified generation, the chromosome with the highest fitness is chosen to represent the recovered task dynamics.

The genetic algorithm has properties that make it particularly attractive in solving the inverse problem. One important property for implementation is that it is very easy to bound the search space and to change the resolution of the parameter search. Also, this procedure is not derivative based, which is important here as analytic derivatives are not available, and function evaluation for numerical evaluation of partial derivatives is expensive. Further, the genetic algorithms have been used in classifier systems, so that the work here can be extended to machine learning.

4. Some computer experiments

The initial tests have been done using speech created by the Haskins articulatory synthesizer, ASY, using task dynamics. In the first series of experiments the conditions of the model used in ASYINV were exactly the same as those that produced the data. Not only was the synthesizer the same, but the details of the task dynamic parameters that were not being recovered were the same, including the articulatory weights [equation (4)] that are a factor in the pseudoinverse calculation. The recovered and original task dynamics are exhibited in Figure 4, in the case when noise with a flat distribution of amplitudes between +- 10 Hz was added to the formant data. This recovery was the best of eight runs of ASYINV, each with an initial population of 100 randomly chosen chromosomes run for 60 generations. Because protruding the lips is has an equivalent effect on the formant frequencies as closing the lips, it was assumed that the lip protrusion target was known. Also, it was assumed that the lips had closed.

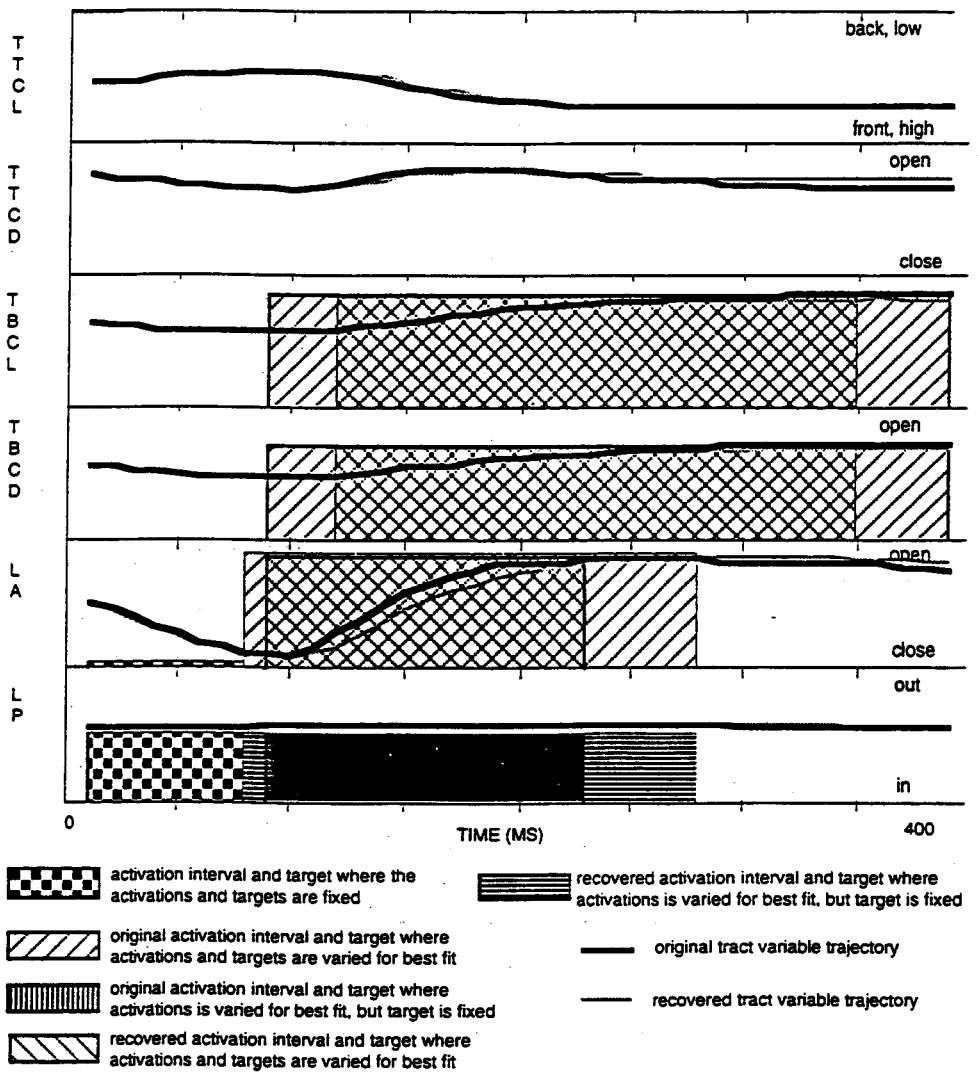

Figure 4: Tract variable trajectories, activation times, and targets for utterance /əbæ/. See Figure 3 for the quantitative limits for each panel.

This was the first attempt at recovery. It can be seen that it did well in recovering the trajectories of the tract variables. That there were substantial errors in the activation intervals did not appear to affect the success in recovering trajectories. While these results were encouraging much more testing was needed, and still needs, to be done. In particular, issues of robustness and efficiency needed to be addressed.

The effect of mismatching the vocal tract that produced the speech data with the one used to recover the task dynamics was tested in further experiments. While the same articulatory synthesizer was used for production and recovery, the articulator weights used in the Jacobian pseudoinverse [equation (4)] were altered from those that produced the data. In these numerical experiments, the utterance that produced the data was a bilabial approximation (LA), which involved the upper lip (ULV), lower lip (LLV), and jaw (JA). Some of the testing involved adding substantially more weight to one of these articulators one at a time. The result was that the recovered articulations produced nearly the same lip aperture (LA) trajectory using the free articulators to compensate for the one that was given extra weight. Thus, for example, with extra weight added to the upper lip, this articulator moved very little, but the lower lip and jaw moved more than in the original, data-producing utterance so that the same total lip aperture trajectory could be achieved. Further work on this can be found in [11].

The function evaluations that are required for the analysis-by-synthesis procedure proposed here is computationally intensive. Something on the order of 1.5 sec. of CPU on a DEC 3000 workstation (Alpha machine) is necessary for one function evaluation for 400 ms of speech. The evolution of a single population of 100 individuals for sixty generations, say for both lips and tongue body involved, can require as many as 2000 function evaluations, for a total CPU time of 50 minutes. For this reason alone, it would be wise to save function evaluations for future use. The saved individuals would specify both the task dynamics and the formant trajectories that go with them. This kind of data base is similar in the field what is known as a *codebook* [12].

In the one attempt at using a codebook, the computer was allowed to "babble" to create an indexed file of 111,596 task dynamic–acoustic pairs. Each individual had a key associated with it denoting the direction and amount of formant transition over a specified interval. (This interval was chosen to be between the release and central portion of the vowel in the data utterances tested.) The amount of formant transition was specified within 10% and, thus, for example, the key IIS201010 denoted an individual with a first formant that increased between 20% and 30%, a second formant that increased between 10 and 20%, and a third formant that was steady within 10%. It was possible to access the individuals based just on the direction of the formants (qualitative match condition), or based both on direction and quantity of change (quantitative matching condition). Given a data utterance, the initial population of individuals used to start the genetic algorithm was chosen in three different ways: randomly, from the indexed data file with qualitative matching, and from the indexed file based on quantitative matching. In the latter two cases individuals were included in the initial population depending whether there was sufficient match between their key and that of the data. If there were more than enough candidates to fill the initial population, then the fittest individuals were chosen to fill the initial population.

Tests were run with initial populations of 100 individuals, run for 60 generations. Each condition was run 8 times. The results for one test utterance, /ədæ/, are shown in Figure

5, where the average maximum fitness for each condition is plotted against generation number.

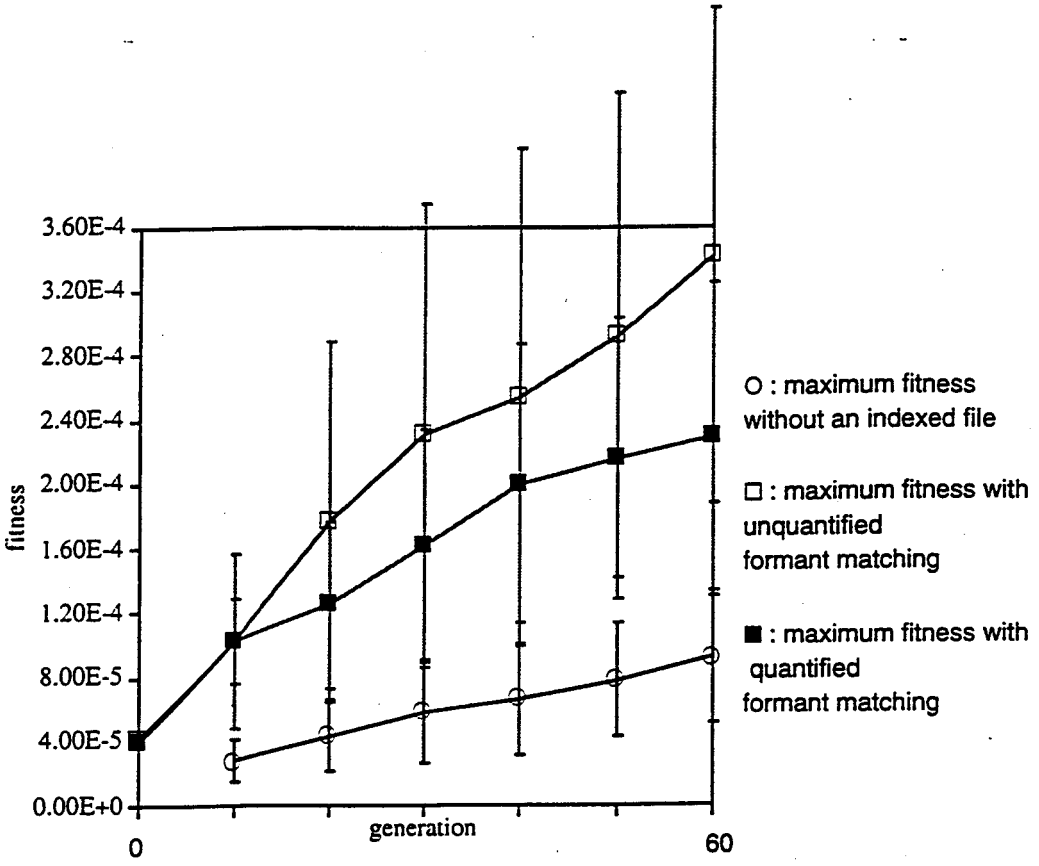

Figure 5: Average maximum fitness plotted against generation number.

This shows that it can help to seed the initial population with likely candidate solutions. However, in comparing the results for populations initialized using quantitative matching versus the results using qualitative matching, it can be seen that being too specific in initializing the population can be detrimental. This is an illustration of the balance that must be struck between search and selection in genetic algorithms. Too much emphasis on taking the fittest individuals initially can mean that the parameter space is not searched thoroughly enough for good optimization. More details on this kind of experiment can be found in [13].

5. Using articulatory context as a constraint in automatic speech recognition

Assume that it is a good thing to use knowledge to reduce variability in the recognition of phonemic (or /CV/ and /VC/) categories. (Without specifying whether phonemes or /CV/s and /VC/s are used as categories, the term *category exemplar* is used to denote a member of a category.) For instance, in deciding the category of a particular vowel sound it is helpful to know whether the speaker is an adult male or a child [e.g. 14]. This knowledge essentially shrinks the category extents defined in an acoustic parameter space, say the first formant versus second formant parameter space. Thus, it

is at least plausible that certain knowledge is useful in reducing variability. We will argue that variability is inherent in speech production, and that some of this variability is the result of context such as that provided by physical mechanisms used to produce speech: the human vocal tracts. Further, it is only through, at least, partial detection and characterization of lawful, contextual covariation that it is possible to form categories and to detect the category exemplars for automatic speech recognition.

Context is always present in speech and it is never possible to factor out context in the sense of producing a prototypical exemplar of a category. Rather, it is necessary each time to find the context and the exemplar of a category simultaneously. Only in this process can one be said to form categories or detect exemplars of categories. Categories exist only in the process of understanding the context. This is the general view that is taken in any methods for automatic speech recognition that are proposed here.

One source of variability for a given speaker is coarticulation. Coarticulation occurs when the production of a category exemplar is affected by the production of neighboring category exemplars. This is always the case in speech, as category exemplars are not produced in isolation in running speech. Thus, a category exemplar can only be defined from samples which are never pure, and is therefore an abstraction from the many instances of occurrence. Further, the problem of coarticulation must be solved (to some extent) simultaneously with the recognition of category exemplars as members of categories: there is no other way. There may be controversy as to which domain to solve the problem, either in the articulatory space or the acoustic parameter space, but reasons for using the articulatory domain will be given later. Ellman and McClellend [15] recognized the importance of exploiting lawful variability caused by coarticulation. They worked with phonetic features and used a neural net that accounted for context to map into the space of phonemes. Shirai and Kobayashi [2,16] have used articulatory recovery as a means to take account of coarticulation, and, thus, proposed to do speech recognition in the articulatory domain.

There are other contextual factors, besides coarticulation, that are provided by the human vocal tract. These include the length of the vocal tract, the wall compliances of the tract, the shape of the hard palate and so forth. While coarticulation affects the physical realization of phonemes produced by an individual according to the context of other phonemes, these other factors depend more on the particular person producing the speech. (Some quantities, such as vocal tract length, are dependent on the speech utterance, so it assumed that default values can be assigned to each individual and deviations from the default can be predicted by context or recovered algorithmically from the speech. For instance, length can refer to the distance along the vocal tract center line from the perpendicular plane containing the teeth to the perpendicular plane containing the cervical vertebra nine.) To recognize category exemplars in a speaker independent way, these other factors must be found.

The use of contextual knowledge is certainly not new to automatic speech recognition. It is usually in the form of "higher-level" syntactic and semantic knowledge that is used to test possibilities that come from the lower level phoneme recognizer [e.g. 17]. We are proposing the serious use of "lower-level" knowledge in the form of articulatory constraint, as have others, most recently in [e.g. 18,19,20]. Both higher and lower level knowledge are intended to remove apparent variability and ambiguity in the category, or symbolic space. All these forms of knowledge use context to do this; the context provided by words, by parts of speech, by phonemes, and by vocal tract anatomy, physiology and physics.

Some have argued that the articualtory domain is the domain to reduce variability caused by intraspeaker context such as that caused by coarticulation and rate. It is now argued that an articulatory domain is the domain that should be used to reduce interspeaker variability context, such as vocal tract size and wall compliance. Working in the acoustic domain may be possible, but it entails implicitly learning rules that are already provided by the physics of the sound production and propagation in the vocal tract. To short circuit this implicit learning, the algorithms that take vocal tract configuration to speech acoustics can be used as an expert system. Further, if there are models of motor control these can also be incorporated as part of the expert system, as with task-dynamics. These physical rules help to constrain the "lower-level" physical system analogous to the way that syntax constrains the "higher-level", linguistic system.

How would the system of articulatory recovery described earlier be used a speech recognition system? One of the first applications that could be considered is speaker adaptation in speech recognition [21]. Before discussing speaker adaptation, simple classifier systems will be discussed.

Task-dynamic parameter recovery was discussed earlier in the paper. Rather than recovering task dynamics on-line, we would propose an off-line, or codebook, technique, where acoustic–task dynamic pairs are stored from previous learning. This would be similar to a stimulus–response classifier systems discussed in the genetic algorithms literature [22,23]. The stimuli would be acoustic parameters derived from the speech signal and the responses would be task-dynamic parameters. Different groups of these stimulus–response pairs would be found for different vocal tract anatomy parameters, so that each group would have a tag specifying such things as vocal tract length and vocal tract wall compliance associated with it. Further, differentiation of the groups could be according to dialect and speaking rate, but these will not be considered here.

Thus, there are two major suppositions that are made here. The first is that an articulatory synthesizer who's dimensions and other physical characteristics can be varied to produce speech of good quality, so that an individual's vocal tract may be mimicked using the synthesizer. Secondly, it will be assumed that for each group of these stimulus–response pairs (a group corresponding to a single vocal tract anatomical specifications), there would be a mapping from the task-dynamic specifications to the symbolic space of category names. This mapping would either be done with expert knowledge (e.g. gestural phonology of Browman and Goldstein[24]) or it would be evolved as a classifier system, just as the acoustic–task dynamic classifiers are evolved. In the latter case the stimulus would be the acoustic–task dynamic specification and the response would be the symbolic representation. The process of training is analogous to the training of a hidden Markov model, i.e. model parameter estimation [25]. Because the mapping from task dynamic–acoustic pairs to the symbolic space is done for each vocal tract type, there would be less variability within sets of such pairs corresponding to a symbol and less overlap between sets corresponding to different symbols.

The procedure for a speaker adaptive speech recognition system would be as follows. The computer itself would generate the acoustic–task dynamic codebook. A standard set of exemplars would be used for different vocal tract dimension settings, with each group of such exemplars representing the utterances spoken by an individual. When the speech of a human user is to be recognized, supervised, test-dependent training might

be used to fit one of these groups to the user [21]. Only a limited set of utterances should be necessary for this, because of the expert knowledge provided by the anatomy of the vocal tract. There are two practical advantages to such a procedure. The first is that training does not have to be done with human speech. Secondly, and also because knowledge of the vocal tract is being used, generalization is relatively easy so that the adaptation phase is brief: A few key utterances should fit a group (i.e. a model vocal tract) to a speaker.

The speech inverse problem continues to be a lively area of research because of its many technical applications [26]. The work here offers a new approach to the inverse problem made possible by knowledge of speech production and, further, is consistent with the trend of providing more constraint in the articulatory domain.

Acknowledgments The work presented here has been supported by the NIH though grant DC-01247 to Haskins Laboratories.

References

[1] J. Schroeter and M.M. Sondhi, "Dynamic programming search of articulatory codebooks," in **Advances in Speech Signal Processing**, ed. by S. Furui and M.M. Sondhi, Marcel Dekker, New York, pp. 231-268 (1992).
[2] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave," *Speech Communication*, **5**, pp. 159-170 (1986).
[3] J.L. Flanagan, K. Ishizaka, and K.L. Shipley, "Signal models for low bit-rate coding of speech", *J. Acoust. Soc. Am.*, **68**, pp. 780-791 (1980).
[4] P. Memelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, **53**, pp. 1070-1082 (1973).
[5] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Am.*, **70**, pp. 321-328 (1981).
[6] R.S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," Speech Communication, **14**, pp. 19-48 (1994).
[7] E.L. Saltzman and K.G. Munhall, "A dynamic approach to gestural patterning in speech production," *Ecological Psychology*, **14**, pp. 333-382 (1989).
[8] L-J. Boë, P. Perrier, and G. Bailly, "The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion," *J. Phonetics*, **20**, pp. 27-38 (1992).
[9] G. Papcun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *J. Acoust. Soc. Am.*, **92**, pp. 688-700.
[10] D.E. Goldberg, **Genetic Algorithms in Search, Optimization, and Machine Learning**, Addison-Wesley Publishing Company, Inc., Reading Massachusetts (1989).
[11] R.S. McGowan and M. Lee, "Task-dynamic recovery of lip approximation using mismatched articulator weights," (submitted).
[12] J. Schroeter, P. Meyer, and S. Parthasarthy, "Evaluation of improved articulatory codebooks and codebook distance measures, in **ICASSP '90**, Albuquerque, NM.
[13] R.S. McGowan, "Recovering task dynamics from formant frequency trajectories: Results using computer 'babbling' to form an indexed data base," in **Feschrift for Katherine Harris**, edited by F. Bell-Berti and L. Raphael, American Institute of Physics, Woodbury, NY (in press).

[14] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **25**, pp. 183-195.
[15] J.L. Ellman and McClellend J.L., "Exploiting lawful variability in the speech wave", in **Invariance and Variability in Speech Processes**, edited by D.H. Klatt and J.S. Perkell, Lawrence Erlbaum Associates, Hillsdale, N.J., pp. 360-385 (1986).
[16] K. Shirai and T. Kobayashi, "Estimation of articulatory motion using neural networks," *Journal of Phonetics*, **19**, pp. 379-385 (1991).
[17] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, S. Seneff, "The Voyager speech understanding system: A progress report," in **Speech Recognition and Understanding**, edited by P. Laface and R. De Mori, Springer-Verlag, pp. 413-424 (1992).
[18] R.C. Rose, J. Schroeter, and M.M. Sondhi, "An investigation of the potential role of speech production models in automatic speech recognition," in **Proceedings of the International Conference on Spoken Language Processing**, Yokohama, Japan (1994).
[19] M.A. Randolph, "Speech analysis based on articulatory behavior," *J. Acoust. Soc. Am.*, **95**, p. 2828 (1994).
[20] G. Ramsay and L. Deng, "A stochastic framework for articulatory speech recognition," *J. Acoust. Soc. Am.*, **95**, p. 2870 (1994).
[21] R. Schwartz and F. Kubala, "Hidden Markov models and speaker adaptation," in **Speech Recognition and Understanding**, edited by P. Laface and R. De Mori, Springer-Verlag, Berlin, pp. 31-58 (1992).
[22] J.H. Holland, K.J. Holyoak, R.E. Nisbett, and P.R. Thagard, **Induction: Processes of Inference, Learning, and Discovery**, Cambridge, Massachusetts: MIT Press (1986).
[23] S. Forrest, B. Javornik, R.E. Smith, and A.S. Perelson, "Using genetic algorithms to explore pattern recognition in immune systems," *Evolutionary Computation*, **1**, pp. 191-211 (1993).
[24] C.P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, **49**, pp. 155-180.
[25] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," The *Bell System Technical Journal*, **62**, pp. 1035-1074 (1983).
[26] J. Schroeter and M.M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. on Speech and Audio Processing*, **1**, PP. 133-150.