In: Bell-Berti, F., & Raphael, L. J. (1995). Producing Speech: Contemporary Issues. For Katherine Safford Harris. AIP Press: New York.

32 Recovering Task Dynamics from Formant Frequency Trajectories: Results Using Computer "Babbling" to form an Indexed Data Base

Richard S. McGowan
Haskins Laboratories

INTRODUCTION

The recovery of vocal tract articulation from speech acoustics, the inverse problem, has been a long-standing problem. Many of the chapters in this book are concerned with the forward problem: How do humans move their articulators to produce the sounds heard as speech? The question here is: Can vocal tract articulation be inferred from the speech signal using machines?

The purpose of the project described in this chapter is to find the extent to which it is possible to recover human articulatory movement from the speech signal. First, it is important to examine the idea of recovering articulation from acoustics. An attempt could be made to recover the area of the vocal tube as a function of distance of a constriction from the lips, but this would tell little of articulation. An articulatory model is needed so that the relations between the area function and aerodynamics and human physiology are defined. What kind of articulatory model is to be used as the domain of the recovered articulation? Is the knowledge required that of the exact path of rigid articulators, such as the jaw, and the exact shape of soft structures, such as the tongue, at every instant? Is even more information sought, such as the muscle activity that causes motion

and shape change? It seems unlikely that such details of articulation will be recovered by a general method in the near future because detailed articulatory models of the entire vocal system do not exist. Even if they did exist, there may be theoretically derivable limits to the amount of detail that can be recovered from the speech signal.

The articulatory models that currently exist contain a relatively low number of degrees of freedom (Coker, 1976; Maeda, 1982; Mermelstein, 1973). Because of the known limitations of these models, such as constraining motion to the midsaggital plane, they can only caricature articulatory movements. In the work described here, the recovery is attempted at a level of description even further removed from the movement of the individual articulators and the details of physiology. This level is the parameter space of task dynamics of the tract variables, known here, simply, as task dynamics (Saltzman, 1986; Saltzman & Kelso, 1987). Task dynamics is instantiated using a present-day articulatory model (Mermelstein, 1973), and it describes the coordinated activity of these model articulators in forming and breaking constrictions in the vocal tract during speech (Saltzman & Munhall, 1989).

It has been argued that a description of a vocal tract based on constrictions is closer to its speech acoustics than a description based on individual articulators (e.g., Boë, Perrier, & Bailly, 1992; McGowan, 1994). Acoustic features, such as resonance frequencies, are more sensitive to the placement and degree of the maximally constricted regions than to any other part of the area function. While task dynamic positions are specified in terms of constriction degree and location, this is not true of an articulatory position specification. Because the task-dynamic level is apparently closer to the resulting acoustics than any of the other articulatory levels discussed, it should be the most readily recoverable from the acoustics among these levels. Also, using task dynamics in the solution of the inverse problem is consistent with the historical trend of using as many constraints as possible to map from the acoustic domain to articulation. For instance, an articulatory model is often employed (e.g., Flanagan, Ishizaka, & Shipley, 1980) instead of an arbitrary area function (e.g., Wakita, 1973) to constrain the articulatory domain. Further, the constraint that articulatory movement be continuous can be imposed. Other constraints arise when the movement of the articulators is assumed to have a particular functional form, such as that of a damped sinusoid (e.g., Parthasarthy & Coker, 1992). By using task dynamics, the movements of the tract variables are constrained to a particular functional form, that of a damped sinusoid.

Recent work on recovering task dynamics from acoustics has shown some promise in model tests (McGowan, 1993, 1994). (It is recommended that the second reference be reviewed by the reader.) These model tests were performed running a program called ASYINV, using formant frequency data produced by an articulatory synthesizer (Rubin, Baer, & Mermelstein, 1981) that incorporates Mermelstein's model vocal tract (Mermelstein, 1973). The data were produced

by writing gestural scores (Browman & Goldstein, 1990) that specify task-dynamic parameters, from which the movement patterns of the articulators in the model vocal tract could be computed. Because each tract variable is assumed to have second-order dynamics, the gestural scores specified, among other things, natural frequency, damping, target position, and the activation intervals over which the these second-order specifications were active. There were constraints imposed to reduce the number of parameters. All systems were assumed to be critically damped, and the duration of activation intervals were always greater than 100 ms and assumed to be equal to the period of oscillation based on the natural frequency. The tract variables used in these tests came in three groups of two, one of each group describing constriction location, and the other constriction degree (Saltzman & Munhall, 1989). One group comprised lip protrusion (LP) and lip aperture (LA); another group, tongue body constriction location (TBCL) and tongue body constriction degree (TBCD); and the third group, tongue tip constriction location (TTCL) and tongue tip constriction degree (TTCD). Each group had the same activation intervals. Thus, within each group only the target specifications for place and degree of constriction were independently specified. The same model vocal tract that produced the data was used in an analysis-by-synthesis procedure to find the optimum task-dynamic parameters--optimum in the sense of obtaining formant frequency trajectories that matched the data as closely as possible in a least-squares sense. The results of these tests encouraged some further research using this approach, and the procedure of testing by using the same models that produced the acoustic data to recover the articulation was retained here.

The recovery program, ASYINV, depends on using constrained models that have been derived from human motor control (task dynamics) and anatomy (Mermelstein's articulatory model). Beyond these biological models, biological analogies can be found in other parts of ASYINV. While the forward problem of speech synthesis is solved using physical law to relate tube shape and aerodynamics to sound, a nondeterministic search procedure is necessary to solve the inverse problem. The program ASYINV employs a genetic algorithm as such a procedure as an alternative for standard optimization procedures. This algorithm is described in a text by Goldberg (1989) and the practical reasons for using this procedure for the inverse problem can be found in McGowan (1994). In this algorithm, the individuals of a population are assigned randomly chosen task-dynamic parameter sets that are coded into binary strings called "chromosomes," and each is assigned a fitness. The fitness used in the present work was the inverse of the sum of square differences between the data formant frequency and the synthesized formant frequency. The sum was taken over the first three formant frequencies in 10 ms intervals for the duration of the utterance. Individuals are chosen to breed with others to form a new population of chromosomes, with the probability of being chosen made equal to each individual's fitness divided by the sum of the fitnesses of the other individuals.

When two individuals mate their chromosomes split at a randomly chosen location with each of two progeny obtaining one part of their chromosome from each parent. The children's fitnesses are evaluated based on their parameter sets as coded in their chromosomes. That is, the task-dynamic model is run based on the parameters specified by each child's chromosome, and the resulting fitness is computed for each child. A small probability of mutation is allowed. Most importantly, this procedure is not strictly an optimization procedure. but it is an adaptive procedure. Given the current population of proposed solutions, it leads to plausible regions in the parameter space that provide better adaptation, even if the fitness function is not stationary. Such an adaptive procedure should be useful if, as Studdert-Kennedy (1991) argues, language development can be "...cast fruitfully in an evolutionary and recapitulatory framework..." (p. 24) with attendant differentiation and assembly for adaptation.

The idea of using biological analogies and models appears reasonable for the inverse problem because it can be argued that human children, especially blind children, in learning to talk must recover articulation from speech. (It will be argued later that children, in some ways, have a less daunting task than recovering the detailed articulation of individual talkers from speech acoustics.) Children are equipped with a mechanism similar enough to other talkers so that they are able to do this (just as ASYINV has a model vocal tract), and in the terminology of learning systems they have an internal model (Holland, 1992; Jordan & Rumelhart, 1992). Not only does the child have a similar mechanism for sound production, but he has plenty of practice through babbling and interaction with others, and babbling sounds have commonalities with adult utterances (e.g., /CV/ constructions). In fact, learning to talk is delayed when a hearing child cannot babble (e.g., Bleile, Stark, & McGowan, 1993). The child also can use experience to help build a mapping between articulation and the resulting sound. While an internal model has been incorporated into the method of articulatory recovery in the previous work, the work to be described here continues the trend of using human-like constraints by allowing "babbling" and access to a data base of previous babbling activity. The quotes are used because the babbling performed here was a random search of the task-dynamic parameter space, while human babbling is more constrained. Also, the term data base is used instead of memory because the formation and access of the computer data base here has little resemblance to human memory (Rose, 1992).

Building a data base from known task dynamic-acoustic pairs was also warranted strictly on practical grounds, because the function evaluations can be computationally costly (say 2 seconds to run task dynamics simulating a 400 ms utterance on a DEC 3000), and a sufficient number of function evaluations needs to be performed to sample the parameter space properly (on the order of 2,000 to 16,000 evaluations in the previous work with a genetic algorithm). It appeared to be prudent to save task dynamic-acoustic pairs from function evaluations (i.e. syntheses) for future use. The task dynamic-acoustic pairs would facilitate

something of a codebook lookup as used by Schroeter, Meyer, and Parthasarthy (1990). That is, the task-dynamic parameter recovery might be enhanced by accessing the data base of function evaluations to help in optimization/adaptation.

In these experiments it was hypothesized that providing the initial population with individuals whose acoustics closely matched that of the data would accelerate the optimization/adaptation. To test this hypothesis, a system for creating a reservoir of task dynamic-acoustic pairs, along with a method for recognizing similarities with the acoustic data needed to be devised. These procedures will now be described, as well as the results of some testing. The emphasis in the testing was to find out whether building a data base of task dynamic-acoustic pairs could help in recovering task dynamics. Measures of the efficiencies of various procedures were not emphasized.

Procedure

The recovery of task-dynamic parameters of four utterances was attempted under various conditions. These utterances were specified by gestural scores for /ədæ/, /əbæ/, /əbi/, and /ədi/ modified to meet constraints imposed by the author, and they are referred to in their unmodified phonetic transcription. The constraints were the same as those in previous work, as described above (see also McGowan, 1994). These scores were used as input to the task-dynamic simulation in order to create four different sets of formant trajectories that served as data in various recovery processes. The genetic algorithm as implemented for the recovery procedure was very close to the Simple Genetic Algorithm described by Goldberg (1989, pp. 59-70), as modified by McGowan (1994). There was only one change to the genetic algorithm from its previous implementation: the binary strings were decoded as Gray code (Forrest, 1993). Gray code ensures that nearest neighbors in an integer representation map to nearest neighbors in a binary representation.

The program used in the previous numerical experiments, ASYINV, was modified so that data structures containing task-dynamic chromosomes and formant frequency trajectories could be stored in an indexed file. These indexed files constituted the data bases in this work. An indexed file allows access to the data structures by means of character-coded keys, here known as acoustic keys. The number of formants tracked, in this case always 3, and the length of the formant tracks in tens of milliseconds, in this case always 39, were recorded in the first 6 characters of the key. The next three characters recorded the direction of change of each formant frequency from 90 ms to 390 ms: increasing, decreasing, or steady. The reason this time interval was chosen was that in all four test utterances the constriction was released at 90 ms and the following vowel extended through 390 ms. A formant frequency was declared steady if changes were less than 10%, and otherwise it was declared increasing or decreasing, depending on the direction of change. The next 6 characters of the

acoustic key quantified the changes in 10% intervals. Thus, the key for /ədæ/ was 390300IDS201010, the key for /əbæ/ was 390300III202010, the key for /əbi/ was 390300III202020 and the key for /ədi/ was also 390300III202020.

Because four different utterances were recovered sharing common indexed files, it was necessary to make the chromosomes a constant length and the decoding of the parameters consistent. Otherwise, the chromosome length and decoding method would have provided extra knowledge as to the values of the task dynamic parameters. In all chromosomes, it was assumed that the lip tract variables LA and LP were activated at most once with the LP target treated as known, that the tongue body tract variables, TBCL and TBCD were activated at most once, and that the tongue tip tract variables TTCL and TTCD were activated at most once. There was a bit attached to each of the groups, lip, tongue body, and tongue tip, that coded whether each of these groups actually was activated. For instance, the gestural score for /əbæ/ did not activate the tongue tip tract variables, and it was possible to express this during recovery in a bit in the chromosome (McGowan, 1993). Also, there were instances where groups were activated more than once in the utterance creating the acoustic data. In these instances one of the activations was treated as known. For instance, the gestural score for /ədi/ had two activations of the tongue tip group, and in the recovery process one of these was assumed to be known. The ranges and resolutions of the parameter values that were covered by the chromosomes are shown in Table 1 (LP is omitted from Table 1 because it is predetermined).

The recovery procedure was run under five different conditions and under each condition, eight times for each utterance. In all of the conditions, the initial population size was 100, and the genetic algorithm was allowed to run for 60 generations with a probability of mating set to 0.6 and the probability of random mutation set to 0.001. The first condition was with the initial population produced randomly without an indexed file, just as in the previous work (McGowan, 1993, 1994), except that the initial population size was larger in this study. The other four conditions included individuals into the initial population from an indexed file derived from a previous session of random babbling.

TABLE 1. *Target value specifications.*

Tract Variable	Maximum/Minimum Target Value	Number of Bits in Chromosome	Resolution
LA	2.33/–0.383 cm	6	0.050 cm
TBCL	3.16/0.51 rad	6	0.042 rad
TBCD	1.63/–0.13 cm	6	0.028 cm
TTCL	1.17/0.40 rad	6	0.012 rad
TTCD	2.15/–0.65 cm	6	0.044 cm

These conditions were the result of combining two different factors: indexed file size and the detail of matching with the data acoustic key required to be taken from the indexed file and included into the initial population. There were two indexed files generated from random babbling. (The same constraints and ranges applied to the task dynamic parameters of the babbled utterances as to those of the original utterances producing the acoustic data.) The small file contained 11,156 individuals and the large file, about ten times as large, contained 111,596 individuals. The other factor that was varied was the number of characters in an individual's acoustic key needed to match the data acoustic key to be included into the initial population, either 9 or 15. With a match of 9 characters in the key, there needed to be agreement in the directions of all three formants, as well as length of utterance and number of formants recorded (and all individuals matched in the two latter parameters). This condition was called the unquantified formant trajectory matching condition. With a match of 15 characters there was the further requirement that the formant trajectories match in the designated 10% intervals. This condition was known as the quantified formant trajectory matching condition. If there were more than 100 individuals in memory that met the matching condition, the 100 individuals that had formants trajectories that matched those of the data best, in the sense of high fitness, were chosen for the initial population. In the cases where there were fewer than 100 individuals that produced the required match, the remaining initial population was filled with randomly generated individuals.

Results

The results of these babbling experiments are reported as the averages of maximum fitnesses found in the populations as a function of generation. The averages are over the eight trials of each utterance without the use of an indexed file, with the use of the small and large indexed file, both with unquantified and quantified formant trajectory matching.

Figures 1(a) and 1(b) show that both large and small indexed files helped increase maximum fitness during the recovery of utterance /ədæ/. In the case of the small indexed file the quantification of the formant trajectory change was not helpful, and in the case of the large indexed file it was actually a hindrance. The large indexed file condition outperformed the small indexed file condition in the final generations for unquantified formant matching. In the four conditions using an indexed file, all initial individuals were drawn from that file, even in the case of a small file requiring a match in quantified formant trajectories (see Table 2). This is unlike what happened for the other utterances.

The other utterances are interesting because their acoustic keys are identical in the first 9 characters (unquantified formant trajectory matching), and in the case of /əbi/ and /ədi/, there was a match in all 15 characters (quantified formant trajectory matching). Those utterances that shared the required number of key

characters drew from the same pool of candidate individuals to be included into the initial population. If there were more than 100 such individuals matching the acoustic keys of two utterances to be recovered, then the initial populations for these two were not necessarily the same, because only the best 100 matches in terms of fitness with the data were taken in this case. However, when there were fewer than 100 matches, the same individuals were taken from the indexed file to be included in the different initial populations. Because there were fewer than 100 individuals taken from memory, /əbi/ and /ədi/ shared the same individuals taken from the small indexed file for both the unquantified and quantified formant trajectory conditions, and the utterances /əbæ/, /əbi/ and /ədi/ shared the same individuals taken from small memory in the unquantified formant trajectory condition.

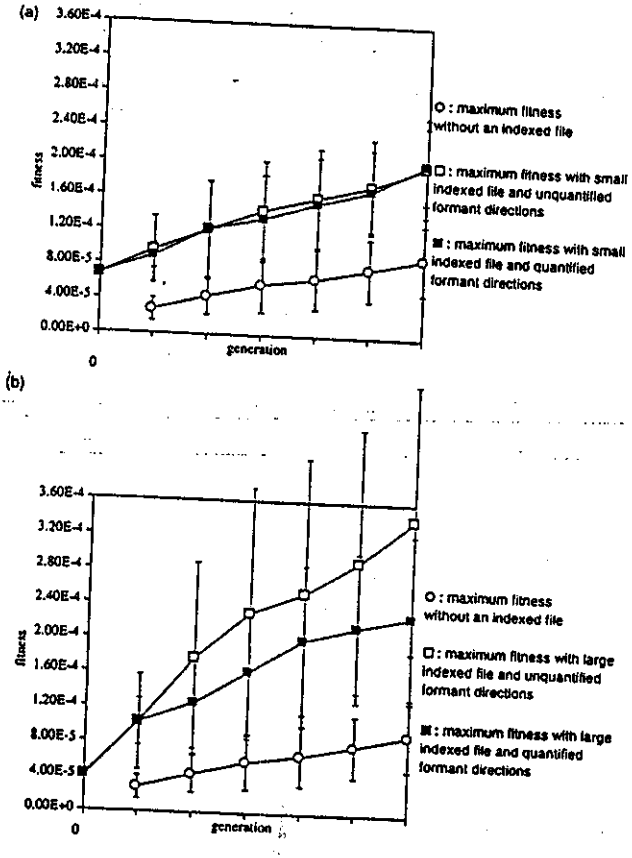

FIGURE 1. (a) Average maximum fitness as a function of generation number for /ədæ/ recovery with the small indexed file; (b) average maximum fitness as a function of generation number for /ədæ/ recovery with the large indexed file.

TABLE 2. *Number of individuals taken from memory.*

Utterance	number taken from the small indexed file for unquantified formant trajectories	number taken from the small indexed file for quantified formant trajectories
/ədæ/	100	100
/əbæ/	51	16
/ədi/	51	15
/ədi/	51	15

Utterance	number taken from the large indexed file for unquantified formant trajectories	number taken from the large indexed file for quantified formant trajectories
/ədæ/	100	100
/əbæ/	100	100
/ədi/	100	100
/ədi/	100	100

The results for /əbæ/ were interesting partly because of the sizable standard deviations (Figures 2(a) and 2(b)). The standard deviations in the fitnesses were on the same order of magnitude as the fitnesses themselves in the cases of no indexed file, and still very substantial for the small indexed file conditions. A detailed look at individual recoveries revealed that there was a tendency in these conditions to sometimes get trapped in local fitness maxima where the tongue tip was used to form the initial consonantal constriction instead of the lips. (For example, this happened 3 out of 8 times for the small indexed file, unquantified formant trajectory matching condition.) The cases where correct initial bilabial closure was attained after 60 generations attained a much higher fitness. Note that Table 2 shows that only 51 individuals were taken from the small indexed file in the unquantified formant trajectory matching condition and 16 in the case of the quantified formant trajectory matching condition. Based on the trials with the small indexed file, it was impossible to tell definitively whether an indexed file may have helped to alleviate the problem of attaining bad local maxima in the fitness function. However, it is obvious that this small indexed file did not expedite the growth of fitness for this utterance on the average. The large indexed file condition helped to sort things out. The unquantified formant trajectory matching condition actually did worse with this particular large indexed file than with the small indexed file, and still with large standard deviations through the generations. Again, 3 out of 8 times the tongue tip was used instead of the lips to form the initial constriction. The quantified formant trajectory matching condition with the large indexed file helped to improve performance above that of all other conditions for /əbæ/. The standard deviations about the average max-

imum fitness were reduced, and there were no instances where the initial bilabial closure was mistaken for a tongue-tip constriction. This was a case where more refined initial characterization of the acoustic signal (i.e., greater acoustic key matching) was helpful. Perhaps this is not surprising, given that /əbæ/ shares the same unquantified formant trajectory key with the utterance /ədi/.

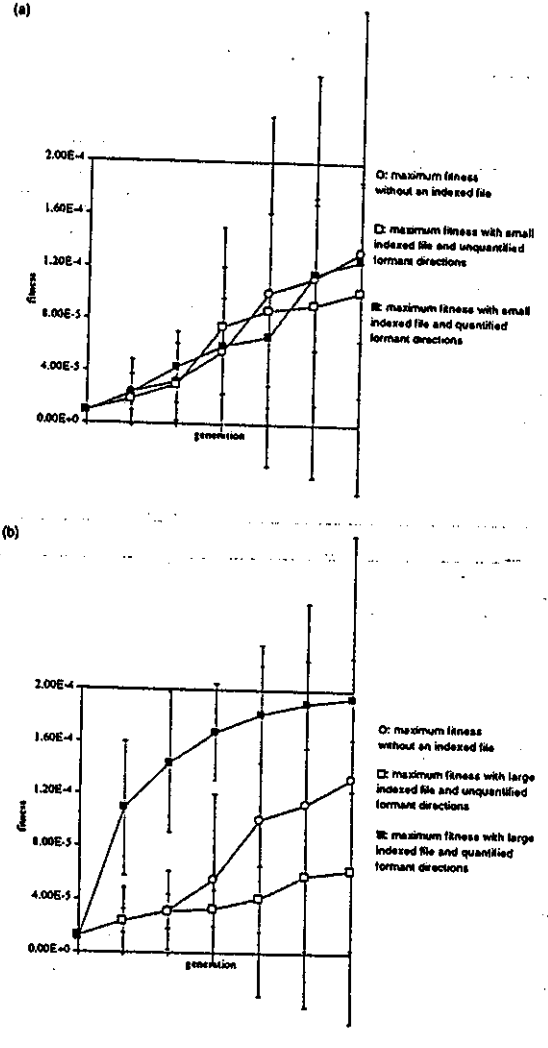

FIGURE 2. (a) Average maximum fitness as a function of generation number for /əbæ/ recovery with the small indexed file; (b) average maximum fitness as a function of generation number for /əbæ/ recovery with the large indexed file.

For the utterance /əbi/ there was clear improvement over the completely random initial population only in the condition of the large indexed file with unquantified formant trajectory matching (Figures 3(a) and 3(b)). This was an instance where the less restrictive, unquantified formant trajectory matching outperformed the quantified formant trajectory matching in the large indexed file condition. Also, despite the fact that this utterance shared the same initial population taken from the small indexed file in the unquantified formant trajectory matching condition with /əbæ/, the problem of substituting tongue-tip constriction for bilabial closure did not result in this condition for this utterance.

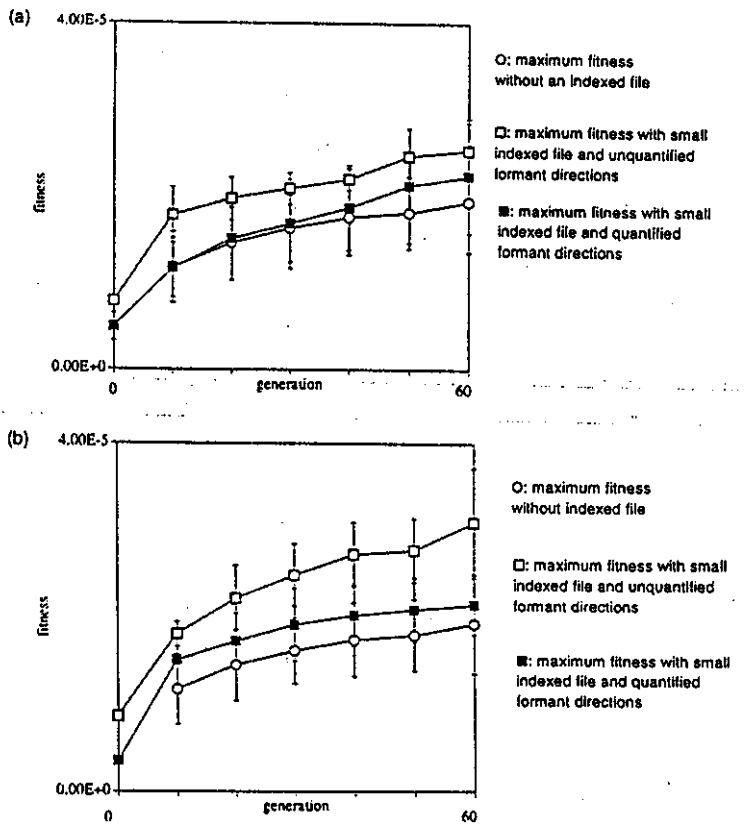

FIGURE 3. (a) Average maximum fitness as a function of generation number for /əbi/ recovery with the small indexed file; (b) average maximum fitness as a function of generation number for /əbi/ recovery with the large indexed file.

The trends for /ədi/ (Figures 4(a) and 4(b)) were very similar to those for /əbi/, with the only clear improvement being in the condition of the large indexed file with unquantified formant trajectory matching. For both cases the maximum fitness as a function of generation was not as great as that for /ədæ/ and /əbæ/, and the speculated reason is that finer control is needed for the tongue body and tip to produce the formants for /i/ than for /æ/. Other reasons were the instability of formant estimation during closure. Also, the lack of information on amplitude of sound and noise content meant that timing of closure release was difficult to estimate.

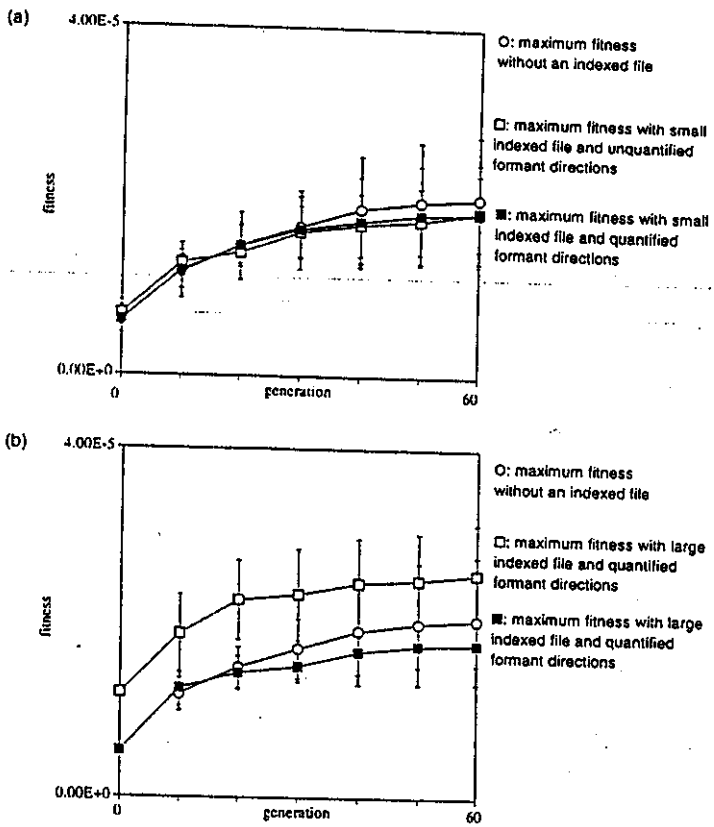

FIGURE 4. (a) Average maximum fitness as a function of generation number for /ədi/ recovery with the small indexed file; (b) average maximum fitness as a function of generation number for /ədi/ recovery with the large indexed file.

CONCLUSIONS

One of the conditions with an indexed file always did better than the condition without an indexed file for all four utterances in terms of average maximum fitness as a function of generation. In the case of /əbæ/, not only was the average of the maximum fitness raised, but the variance over the trials was decreased, so there was more consistency in the better result. The larger indexed file always seemed to outperform the smaller indexed file, but there was no consistency as to which key matching condition did the best. To generate that large indexed file took some computing effort, and may only be justified if a large number of utterances is to be recovered (at 16,000 function evaluations per utterance without memory, at least 7 recoveries with memory should be planned in order to break even). The inconsistency in the efficacy of different matching conditions remains to be explored. There appears to be a balance that needs to be struck between specificity and variety in the initial population. Obviously, if only individuals with correct behaviors are chosen with a highly specific matching condition, then variety is not an issue, but there is no guarantee that this will be the case. /ədi/ and /əbi/ had the same quantified formant trajectory keys despite the fact that the initial places of articulation were different. Perhaps because of this, and the fact that incorrect places of articulation can appear to provide a relatively good fit to the data in a new population, these utterances required an initial population with more variety than could be obtained from matching quantified formant trajectory keys.

There would be better results in all conditions if the number of unknown task dynamic parameters was reduced. There were 41 bits in each chromosome, which resulted in a search space of about 2.2×10^{12} possibilities. There may be ways of reducing the search space size that are more effective than others. For instance, for /əbæ/, to know that the lips formed the initial closure would greatly help the algorithm along. When measurements of articulatory movements are available they should be used to reduce the unknowns, so that other articulatory movement can be inferred directly. Aerodynamic information would be useful in obtaining timing of releases.

These model tests provide a means of determining how much detail can be recovered from an acoustic signal. For instance, it may be the case that task-dynamic parameters are recoverable, but the movements of individual articulators are not. The recovery of individual articulators was not addressed in this paper. However, if this is the goal in the future, the model vocal tract needs a sufficient number of degrees of freedom to move as an exact mimic of a person who creates speech. And, if the model is anatomically realistic, the number of degrees of freedom could grow to be very large. Also, to recover articulatory movement from speech acoustics using task dynamics requires that parameters called articulatory weightings be recovered (Saltzman & Munhall, 1989). These weights determine how much of each articulator is used in

attaining a task-dynamic goal. For example, a bilabial closure can be attained with little jaw movement if the lips move sufficiently. The extra degrees of freedom, derived from a realistic articulatory model, and the extra unknown parameters provided by articulatory weights, may mean that there is no robust procedure for recovering the detailed movement of each articulator. This may be possible only when the acoustics is measured simultaneously with some articulatory movements.

An easier task than complete recovery is teaching a machine to repeat speech intelligibly, assuming that it has some internal model of vocal tract anatomy and motor coordination (e.g., task dynamics). A human could train a machine to utter intelligible utterances based on its own internal models. The human trainer would rate the machine on how well it thought it did, and this allows for a necessary subjective element in the fitness function. This fitness function will not be perfectly repeatable, so that the genetic algorithm as an adaptation, rather than as an optimization algorithm, would be best suited for the task: as the trainer changes his impressions, so do the utterances change. The fitness function evolves as the human-machine partnership evolves, thus resembling an endogenous fitness function (Holland, 1992, 1993). In this work, data bases have been used with a nonassociative reinforcement learning algorithm to expedite the inverse solution (see Barto, 1992, for associative and nonassociative learning). Instead of imposing acoustic keys to use to choose the initial population, an associative learning system should be employed so that the acoustic keys evolve with a human trainer along with the correct maps between acoustics and task-dynamics. A classifier system could provide such a learning algorithm (Holland, 1992).

Is this closer to what children are required to do in learning speech, rather than recovering the detailed articulation of speakers? They perform a distal learning task with adult teachers judging the acoustical output as the children learn to speak (Jordan & Rumelhart, 1992). While the children have a physically instantiated internal forward model--a vocal tract with resultant speech--they do not use it to produce exact acoustic mimics of adult utterances. If children learn an articulatory-acoustic mapping for speech in general, they must have another internal model (an uninstantiated articulatory-acoustic mapping) that allows for differences in the anatomy and motor control biases of different speakers. It is for the adults to teach the children what in the acoustic signal is worth their attention. As stated by Jordan and Rumelhart (1992): "...the role of external 'teachers' is to help with... representational issues rather than to provide proximal targets directly to the learner" (p. 346). Thus, the adult may respond favorably to an utterance that was produced with a task dynamics that is appropriate and transcends the problem of anatomical scaling. This would indicate that such an adult knows the mapping between task dynamics and the speech acoustic signal. For the child, the physical vocal tract provides a tool in the construction of this more abstract acoustic mapping.

Could it be that children learn to talk by learning the mapping between task dynamics and the acoustics of speech? The task dynamic domain provides flexibility in relation to articulation, while its relation to the acoustic domain is sufficiently well determined. There must be sufficient flexibility in the articulatory domain, because vocal tracts vary anatomically among people. However, to recognize a particular gesture, or sound production as a particular meaningful utterance the acoustic output must be of a certain pattern. Task dynamics that map into particular acoustic features slowly mature as the child becomes an adult. As far as movement is concerned, the things that are targeted could be the task dynamics, because these parameters are acoustically the most salient. The articulation that achieves the task dynamics specification is assembled according to individual anatomy and motor control biases. These are plausible reasons for the articulatory-acoustic mapping to be learned through task dynamics; but this has yet to be shown to be the case.

ACKNOWLEDGMENTS

This work was supported by NIH Grant DC 01247 to Haskins Laboratories. I wish to thank Kathy Harris for helping me get started in speech research.

REFERENCES

Barto, A. G. (1992). Reinforcement learning and adaptive critic methods. In D. A. White & D. A. Sofge (Eds.), *Handbook of intelligent control* (pp. 469-91). New York: Van Nostrand Reinhold.

Bleile, K. M., Stark, R. E., & McGowan, J. S. (1993). Speech development in a child after decannulation: Further evidence that babbling facilitates later speech development. *Clinical Linguistics & Phonetics, 7*, 319-337.

Boë, L.-J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics, 20*, 27-38.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299-320.

Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE, 64*, 452-460.

Flanagan, J. L., Ishizaka, K., & Shipley, K. L. (1980). Signal models for low bit-rate coding of speech. *Journal of the Acoustical Society of America, 68*, 780-791.

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

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley Publishing Company.

Holland, J. H. (1992). *Adaptation in natural and artificial systems.* Cambridge, MA: MIT Press.

Holland, J. H. (1993). *Echoing emergence: objectives, rough definitions, and speculations for echo-class models* (No. 93-04-023). Santa Fe Institute.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16,* 307-354.

Maeda, S. (1982). A digital simulation model of the vocal tract system. *Speech Communication, 1,* 199-229.

McGowan, R. S. (1993). Implementing a genetic algorithm to recover task-dynamic parameters of an articulatory synthesizer. *Haskins Laboratories Status Report on Speech Research, SR-113,* 95-106.

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

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53,* 1070-1082.

Parthasarthy, S., & Coker, C. H. (1992). On automatic estimation of articulatory parameters in a text-to-speech system. *Computer speech and language, 6,* 37-75.

Rose, S. P. R. (1992). *The making of memory: from molecules to mind.* New York: Doubleday.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70,* 1109-1121.

Saltzman, E. L. (1986). Task-dynamic coordination of speech articulators: A preliminary model. *Experimental Brain Research, 15,* 129-144.

Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled action: A task-dynamic approach. *Psychological Review, 94.* 84-106.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamic approach to gestural patterning in speech production. *Ecological Psychology, 14,* 333-382.

Schroeter, J., Meyer, P., & Parthasarthy, S. (1990). Evaluation of improved articulatory codebooks and codebook distance measures. *ICASSP '90.* Albuquerque.

Studdert-Kennedy, M. (1991). Language development from an evolutionary perspective. In N. A. Krasnegor, D. M. Rumbaugh, R. L. Schiefelbush, & M. Studdert-Kennedy (Eds.), *Biological and behavioral determinants of language development* (pp. 5-28). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Wakita, H. (1973). Direct estimation of vocal tract shape by inverse filtering of acoustic speech wave forms. *IEEE Trans. Audio and Electroacoustics, 21,* 417-27.