



ELSEVIER

Speech Communication 14 (1994) 19-48

900

SPEECH
COMMUNICATION

Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests

Richard S. McGowan

Haskins Laboratories, 270 Crown Street, New Haven, CT 06511, USA

(Received 8 December 1992; revised 2 August 1993)

Abstract

Articulatory trajectories of an articulatory model were recovered by means of a genetic algorithm from acoustic information using a task-dynamic model of speech articulation. Tests on simulated utterances /əbæ/ and /ədæ/ show that the method can recover most of parts of an original trajectory, but it has trouble in obtaining precise timing. For the recovery of articulation, formant frequency trajectories should be supplemented by additional acoustic information, such as RMS amplitude.

Zusammenfassung

Auf der Grundlage eines aufgaben-dynamischen Modells wurde mit Hilfe eines genetischen Algorithmus die Artikulationstrajektorien eines Artikulationsmodells aus akustischer Information rekonstruiert. Tests mit einer simulierten Äußerung /əbæ/ und /ədæ/ zeigen, daß die Methode einen Großteil der ursprünglichen Trajektorien rekonstruieren kann. Schwierigkeiten treten jedoch in der exakten zeitlichen Koordination auf. Zur Rekonstruktion von Artikulation sollte zusätzliche akustische Information, wie zum Beispiel die RMS Amplitude, die Trajektorien der Formantenfrequenzen ergänzen.

Résumé

A partir de l'information acoustique, il a été possible, avec un algorithme génétique, de retrouver des trajectoires articulatoires, en se servant d'un modèle dynamique de production de la parole. Des tests sur des logatomes simulés /əbæ/ et /ədæ/ montrent que cette méthode peut recouvrir la plupart des trajectoires originales; cependant le traitement du timing précis pose des problèmes. Pour récupérer l'articulation, il est nécessaire de rajouter, aux trajectoires de fréquences formantiques, de l'information acoustique supplémentaire, comme l'amplitude RMS.

Key words: Articulatory recovery; Inverse problem; Task dynamics

1. Introduction

“We suggest, however, that a recognizer using the anatomical or neurophysiological rather than the acoustical level of representation at this stage would more nearly simulate the process of human speech reception.” (Stevens, 1960.)

Since this statement was written, many tools have been developed so that the goal of mapping from acoustic information to articulation may now be practical. Articulatory synthesis has slowly developed over the last forty years producing more realistic articulatory models and speech output (Dunn, 1950; Stevens and House, 1955, Mermelstein, 1973; Coker, 1976; Maeda, 1982; Liljencrants, 1985; Sondhi and Schroeter, 1987; Lin, 1990). The study of skilled activity has developed concepts and models, such as the task-dynamic model of speech articulation, that allow an account of how individual articulatory movements are organized for speech tasks (Saltzman and Munhall, 1989). Linguists have developed phonological systems based on articulatory movement that use task dynamics in their computational implementation (Browman and Goldstein, 1990). Finally, the advances in computer technology, even in the last couple of years, have allowed the solution of optimization problems that require large numbers of complicated function evaluations on relatively inexpensive machines in a reasonable time. Thus, stochastic methods, such as genetic algorithms (Goldberg, 1989), can be tested on the problem of finding optimal maps from acoustics to articulation.

The recovery of articulatory movement from speech acoustics, that is the solution of the speech inverse problem, has been a subject of research for technological applications: automatic speech recognition, low bit-rate coding and text-to-speech synthesis (see (Schroeter and Sondhi, 1992) for a review, particularly for low bit-rate coding). Some early work on articulatory recovery used linear predictive coding to map from the acoustic domain to area functions, e.g. (Atal and Hanauer, 1971, Appendix F; Wakita, 1973). Attempts to map from limited acoustic information, such as the first three formant frequencies, to the area

function of a vocal tract are not always successful because many area functions can produce identical, or nearly identical acoustic data (Mermelstein, 1967; Schroeder, 1967). An articulatory model that constrains the area function in ways that the human vocal tract constrains the area function may help to simplify this one-to-many problem. Also, the use of an articulatory model can be helpful for low bit-rate coding and text-to-speech synthesis because of the relatively low speed with which the vocal tract articulators move. For speech recognition problems there is the possibility of factoring out coarticulatory effects with the further addition of a good model of articulatory control. However, even with a low-dimensional articulatory model, Atal et al. (1978) found that the logarithms of the first three formant frequencies were not enough to specify the model's articulatory settings. Flanagan et al. (1980) used an articulatory model and the squared difference of log amplitude spectra as an error measure in their optimization for vocal tract shape recovery. They mentioned the use of temporal continuity constraints applied to the articulators during a series of optimizations of continuous speech as a means to avoid ambiguity in the acoustic-to-articulatory map. Interestingly, they reduced the number of articulatory parameters to be recovered from the acoustic signal by measuring the mouth opening directly. However, it was necessary to include subglottal pressure as an articulator because of its effect on the output speech amplitude. Levinson and Schmidt (1983) used the articulatory model of Coker (1976) and a spectral difference error to map acoustics into articulation. They had trouble with the ventriloquist effect (the one-to-many problem), although it is not clear whether a continuity constraint in time would have completely addressed the problem that emerged in their investigation.

Still more constraints have been applied in the articulatory domain by using models for articulatory movement. Shirai and Kobayashi (1986) described an optimization method using various spectral and cepstral error measures. Because they were interested in speech recognition, they parameterized the articulators as second-order systems with step input “commands” that were

recovered in the optimization. By recovering underlying control they were able to remove coarticulatory effects. Meyer et al. (1989) used a Kalman filter with articulatory constraints to map from an ARMA model of the acoustic pressure wave to an articulatory model. In a text-to-speech application, Parthasarthy and Coker (1992) mapped from the acoustic domain to an articulatory model at the phone level. For each phone there were targets, transition times and transition speeds of a specified functional form for each of the articulatory coordinates, whose values were context dependent. These investigators constrained the articulatory domain beyond that of continuity in time to specific functional forms for the articulatory movements.

Further developments have been made in the use of codebook look-up (Schroeter et al., 1987; Larar et al., 1988) and the use of neural networks, e.g. (Shirai and Kobayashi, 1991; Papçun et al., 1992; Rahim et al., 1993). These advances are classifiable as off-line techniques, because articulatory-acoustic correspondences are either stored explicitly, or they are learned by a neural net. For codebook look-up, articulatory-acoustic correspondences are generated to form a data base. To perform an acoustic-to-articulatory mapping, the codebook is accessed for possible starting points for further optimization. This access to good starting conditions makes it easier to avoid local minima. Again, an articulatory continuity constraint in time can be used to avoid ambiguity in the case of running speech, but improved access beyond a simple continuity constraint can be provided by using dynamic programming (Schroeter and Sondhi, 1989). This procedure is more constraining than just continuity in time, because the entire utterance is considered in the optimization for recovering the corresponding series of vocal tract shapes. Random sampling of the articulatory space, with controlled pruning of articulatory configurations, provides a means of covering the acoustic space for codebook generation (Schroeter et al., 1990). Neural nets have the potential advantage of reducing computation times and storage requirements over those of codebooks. The work of Papçun et al. (1992) used human movement and acoustic data to train neu-

ral nets. They found that there were critical articulators depending on the kind of utterance, for example the tongue tip for an alveolar release. These critical articulators had less variability in their movement patterns than did the other articulators, and as a result the neural nets did a better job of tracking their movements. Techniques used in codebook access, such as dynamic programming, have been used to improve the performance of neural nets (Rahim et al., 1993).

The intent of the present work, a preliminary part of which is reported here, is to study the amount of articulatory movement that could conceivably be recovered from acoustic speech data in a speech physiology and production laboratory setting. To recover the movement of all the articulators simultaneously may prove to be impractical in most real-world conditions. However, in a laboratory situation, where some movement can be measured directly, as with photography or magnetometers (Flanagan et al., 1980; Perkell et al., 1992), a recovery technique may prove to be practical and useful. Thus, the goal sought here is different from the ones cited so far. While the previous results in these areas bear consideration for the current task, a means of automatic-speech recognition, low bit-rate coding and text-to-speech synthesis is not offered in this work. Also, a model of human speech recognition, as highlighted in the leading quote from Stevens, is not offered here either, although theories of speech recognition have taken articulatory recovery as essential for this process, e.g. (Stevens, 1960; Liberman and Mattingly, 1985). Among other laboratory techniques for extracting articulator configurations from sound, there has been some previous work involving the use of impedance tubes with externally generated sound sources (e.g. (Sondhi and Resnick, 1983; Milenkovic, 1987)), and two-point pressure measurements, with one point near the glottis, (Milenkovic, 1984). The goal here is slightly different from the goal of those studies, however, in that only speech acoustic data is to be used. These previous methods and the one proposed here are not exclusive of one another and may, in fact, be complementary.

The method envisioned for articulatory recovery is analysis-by-synthesis, much as those of pre-

vious works cited above. As the articulators move during speech, pressure waves are recorded, and acoustic parameters are extracted from these data. To recover the movements of the articulators an optimization algorithm is used to adjust the articulator trajectories of an articulatory synthesizer so that the parameters of its acoustic output match those of the original data.

This report is restricted to the results of model experiments, where an articulatory synthesizer is used to produce the data and the same synthesizer is used to recover its own movements. Thus, since all the generating principles are explicitly known, this is a best case study of the limitations that could be encountered in attempting a mapping back from acoustics to articulation using the method proposed. The method includes, among other things, particular acoustic parameters, a particular articulatory model and synthesizer, and an optimization algorithm with an error measure. These model tests provide a means for tuning the method with additional acoustic information or improvements to the articulatory model, and, perhaps, with modifications to the optimization procedure and error measure. The difficulties that could be encountered in progressing from model studies to actual measurement will be detailed as the method is described.

In brief, the acoustic data are restricted to the trajectories of the first three formant frequencies and the optimization technique used for this series of numerical experiments is provided by a genetic algorithm (Goldberg, 1989). The task-dynamic model developed by Saltzman, Kelso and others (Saltzman, 1986; Saltzman and Kelso, 1987; Saltzman and Munhall, 1989) is used to produce articulatory trajectories. The task-dynamic model uses a set of dynamic parameters and geometric transformations to be described in the next section to simulate the movements of the articulators. It is the dynamic parameters of the task dynamics that are varied to produce the optimal match in the formant trajectories.

2. Method

The articulatory synthesizer used in these experiments was the Haskins Laboratories articula-

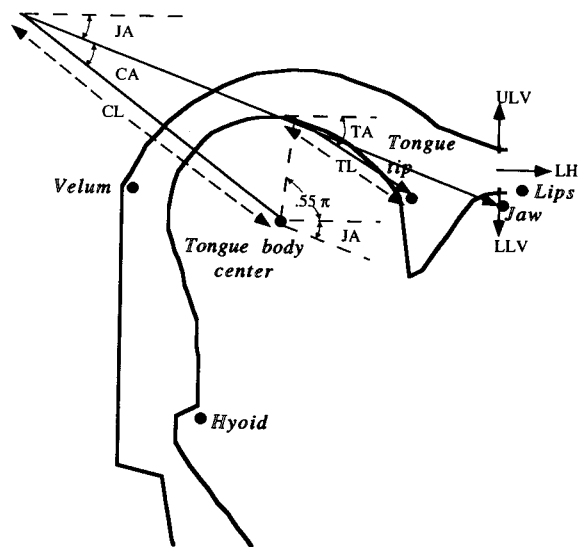


Fig. 1. Model vocal tract with articulators used in the Haskins Articulatory Synthesizer: ASY.

tory synthesizer, ASY, as described by Rubin et al. (1981), which uses Mermelstein's articulatory model (Mermelstein, 1973). The articulators, whose positions can vary and thus change the shape of the vocal tract, are shown in Fig. 1. The tongue body center, tongue tip, jaw and lips were the articulators used in the tests reported here. The coordinates of the jaw are specified by the angle of a fixed-length vector, JA, with origin at the condyle and end at the point marked jaw in Fig. 1. The position of the tongue body center is specified by a vector relative to the jaw vector, with its origin at the condyle and end at the tongue-body center articulator. This vector has a length, CL, and an angle, CA, relative to the jaw. The tongue tip is specified by a vector of length TL whose origin lies on the outline of the tongue body and whose angle relative to the jaw and tongue body is TA. The lips are specified in Cartesian coordinates, with the vertical dimension specifying the dimension of lip closure/opening, and the horizontal specifying protrusion. The upper lip's vertical position, ULV, is specified in relation to the fixed skull, and the lower lip vertical position, LLV, is specified in relation to the jaw. The lips are yoked in the horizontal

dimension, so this coordinate is specified for both lips as lip horizontal, LH.

The articulatory positions can be specified by a table of coordinate values, where each row of the table corresponds to a time in the movement trajectory. In the cases reported here, the positions of the articulators were specified and a rational transfer function was calculated every 10 ms, according to the Kelly–Lochbaum algorithm (Kelly and Lochbaum, 1962), see also (Rubin et al., 1981). Formant frequencies were obtained from each transfer function by applying a Fourier transform to the denominator of the transfer function and using a peak-picking algorithm on the resulting magnitude spectrum. Thus, formant frequency trajectories were created with a 10 ms frame rate for both the data file and the proposed articulatory solutions during the optimization. To synthesize the speech was unnecessary for these model experiments. In actual applications of this method, speech pressure waves would have to be analyzed and measured formant values placed in a data file. Also, synthetic speech generated from proposed articulatory solutions will have to be examined as additional parameters, such as RMS amplitude, become incorporated into the acoustic parameter list.

The measure of error between the data and the proposed solution was the sum, taken over 10 ms intervals, of the squares of the difference between the formant frequencies of a proposed solution and those of the speech data, for the first three formant frequencies.

$$\text{error} = \sum_{i=1}^3 \sum_{j=1}^N (f_{ij}^{\text{model}} - f_{ij}^{\text{data}})^2, \quad (1)$$

where f represents formant frequency, the i subscript denotes the formant number, the j subscript represents the 10 ms frame number, and N is the total number of frames. In fact, the inverse of this error measure was used as a measure of fitness for the genetic algorithm, as will be described later.

$$\text{fitness} = 1/\text{error}. \quad (2)$$

The error in Eq. (1) is an unsophisticated error measure, as the more recent literature shows, e.g.

(Shirai and Kobayashi, 1986). This simple choice of error function allowed an evaluation of how well the method does with an acoustic data set known to be impoverished, e.g. (Mermelstein, 1967). This error measure may not be the best choice in terms of avoiding local minima in an optimization, but this can be evaluated at a later date. Further, the choice of error measure should be guided by the application in question. For instance, in the applications of low bit-rate coding and text-to-speech synthesis the error measure should be sensitive to perceptual mismatches (Parthasarthy and Coker, 1992; Schroeter and Sondhi, 1992). Ultimately, the error measure for the current purpose will have to be guided by how well articulation is tracked, and more a sophisticated error measure will undoubtedly result, particularly if laryngeal parameters are to be recovered (Flanagan et al., 1980; Schroeter et al., 1987).

Shirai and Kobayashi (1986), Parthasarthy and Coker (1992) and McGowan (1991) proposed that articulatory trajectories be recovered as parameterized functions over /CV/ or /VC/ intervals. These procedures help constrain the articulatory domain enough so that the one-to-many problem in mapping from acoustics to articulation is alleviated, similar to the way continuity constraints help. There are two, not necessarily exclusive, ways that the one-to-many problem could occur. First, a connected region of acoustic parameters could correspond to unconnected regions in the articulatory space, so that there is no path in the articulatory parameter space that connects these different regions along which the corresponding acoustics remains constant, or nearly so. If this is a problem, it could be handled easily by continuity constraints over time. The other way that nonuniqueness manifests itself is when a small region of acoustic parameters corresponds with a large connected region of the articulatory parameter space, e.g. (Atal et al., 1978). Linearizing the articulatory-to-acoustic, or forward, mapping would result in a mapping with singular values equal to zero. Practically the problem is close to the problem of sensitivity, because singular values near zero in the forward mapping indicate sensitivity in the inverse mapping, in that a small

change in the acoustic data can produce a large change in the recovered articulation. Because singular values are not calculated exactly, sensitivity could be an indication of a many-to-one problem. Continuity constraints or constraining the problem further by assuming that articulatory trajectories are parameterized functions over time should diminish these problems.

2.1. Task dynamics

To constrain the articulatory model in the present work, the task dynamic model (Saltzman, 1986; Saltzman and Kelso, 1987) was employed. Rather than parameterizing the articulator position trajectories themselves, the task-dynamic model was allowed to drive what are known as tract variables. These are variables describing constriction locations and degrees. Of particular interest in this work were the tract variables tongue body constriction location and degree,

TBCL and TBCD, tongue tip constriction location and degree, TTCL and TTCD, lip protrusion, LP, and lip aperture, LA. With task dynamics applied to the vocal tract (Saltzman and Munhall, 1989), tract variables are given a dynamics that will be described below. The two sets of coordinates, the tract variables and the ASY articulator coordinates, are related to each other by geometric transformations to be discussed below. Fig. 2 illustrates the tract variables' relation to the ASY vocal tract and tabulates some of the tract variables' dependence on articulatory variables of ASY. Thus, with the task-dynamics of the tract variables specified, and with the geometric transformations between the tract variables and the ASY articulators also specified, the articulator trajectories in time are determined, and the formant trajectories can be computed.

The geometric transformations between tract variables and articulators are derived based on Mermelstein's articulatory model (Mermelstein, 1973). For instance, the tract variable lip aperture, LA, depends on the articulatory coordinates upper lip vertical, ULV, lower lip vertical, LLV, and jaw angle, JA. The mathematical relation is

$$LA = (\text{vertical position of upper teeth} + ULV) - (\text{length of jaw vector} \cdot \sin(JA) + LLV). \quad (3)$$

		ASY ARTICULATOR COORDINATES ($\theta_j, j = 1, 2, \dots, n; n=8$)							
TRACT VARIABLES ($Z_i, i = 1, 2, \dots, m; m=6$)		LH (θ_1)	JA (θ_2)	ULV (θ_3)	LLV (θ_4)	CL (θ_5)	CA (θ_6)	TL (θ_7)	TA (θ_8)
LP (Z_1)		●							
LA (Z_2)			●	●	●				
TBCL (Z_3)			●			●	●		
TBCD (Z_4)			●			●	●		
TTCL (Z_5)			●			●	●	●	●
TTCD (Z_6)			●			●	●	●	●

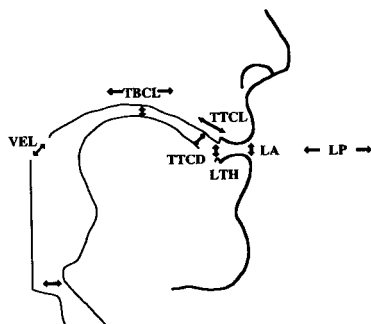


Fig. 2. Tract variables and their dependence on ASY articulators.

The vertical position of the upper teeth and the length of the jaw vector are fixed parameters. Recall that ULV is measured relative to the fixed upper teeth and LLV is measured relative to the position of the end of the jaw. The tract variable LP is directly proportional to the articulatory coordinate, horizontal lip position, LH. The tract variables TBCL, TBCD, TTCL and TTCD are measured in a head-centered polar coordinate system, whose origin is 3.85 cm below the point on the outline of the palate that has the maximum vertical value (Fig. 2) (see (Mermelstein, 1973) for details). This puts the origin of the head-centered system 7.34 cm anterior and 4.56 cm inferior to the condyle (origin of the jaw vector in Fig. 1), where it also provides the center for the circular part of the outline of the upper/rear wall of the model vocal tract. Angles

are measured from the horizontal, with positive angles counter clockwise in orientation. The tract variables TBCL and TTCL give the angular positions of the maximum constriction formed by the tongue body and tongue tip, respectively. TBCD and TTCD are the distances from the fixed upper/rear vocal tract wall to the tongue body and tongue tip, respectively, at the positions of maximum constriction.

In abstract form the transformation from the articulatory coordinates to the tract variables can be written as (Saltzman and Munhall, 1989)

$$z = z(\phi), \quad (4)$$

where z is a vector of the tract variable positions and ϕ is a vector of articulator positions.

The task dynamics of the tract variables is described by a set of uncoupled, linear, second-order equations. With the task-dynamic parameters of mass, damping coefficient and stiffness of each tract variable specified in the diagonal matrices M , B and K , respectively, and target positions for the tract variables specified in the vector z_0 :

$$M\ddot{z} + B\dot{z} + K(z - z_0) = 0. \quad (5)$$

(Here, it will be assumed that M is the identity matrix. Normalizing each equation by the mass does not affect the possible solution set because the coefficient matrices are diagonal.) Further, the time intervals for which the target position of a tract variable is other than the default rest position are specified. These time intervals, known as activation intervals, are specified by starting and ending times. The mass, damping coefficient, stiffness and target position of a tract variable are constant during each of its activation intervals.

Tract variables recruit various articulators as their dynamics are instantiated in the vocal tract. The set of task dynamics equations transformed into the articulatory space becomes a coupled set of equations because the Jacobian of the z transformation of Eq. (4) is not diagonal. From Eqs. (4) and (5):

$$M(J\ddot{\phi} + \dot{J}\dot{\phi}) + BJ\dot{\phi} + K(z(\phi) - z_0(\phi_0)) = 0, \quad (6)$$

where J is the Jacobian matrix for the z transformation, i.e. the i - j element of J is the partial derivative of the i -component of z with respect to the j -component of ϕ . Also, $z_0 = z(\phi_0)$. Solving for the acceleration of ϕ :

$$\ddot{\phi} = J^* \left\{ (M^{-1}[-BJ\dot{\phi} - K\Delta z(\phi)]) - \dot{J}\dot{\phi} \right\}, \quad (7)$$

where $\Delta z = z - z_0$ and J^* is a weighted Jacobian pseudoinverse. The articulators are recruited at various weightings to achieve the target as specified by task dynamics at the tract variable level and these weightings must be given to complete the task-dynamic specification. The importance of each articulator in the activation of a tract variable is expressed by an articulatory weight in J^* (see (Saltzman and Munhall, 1989, Appendix 2) for details). Also, there are other details about task dynamics that are discussed in (Saltzman and Munhall, 1989), including the return to a neutral target after activation, and how different tract variables can use the same articulator simultaneously.

Eq. (7), which represents a set of coupled, nonlinear differential equations, is solved by a numerical extrapolation, in this case, a fourth-order Runge–Kutta time stepping routine. This creates articulatory trajectories. Note, however, it is the set of task-dynamic parameters contained in M , B , K and z_0 , as well as the activation times and articulatory weightings, that determine the evolution of both the tract variables and articulators. Once the task-dynamic parameters are specified, the articulator position trajectories are specified by this set of coupled differential equations. Once the articulators' movements are obtained, formant trajectories can be computed from the ASY transfer function.

To recover the articulatory movement in ASY, the task-dynamic parameters were adjusted so that the resulting formant trajectories fit the data trajectories. Without constraints, the task-dynamic parameters that would have been needed to be adjusted were the task-dynamic parameters: M , B , K and z_0 , as well as the activation times and articulatory weightings. Assuming that M was the identity, the optimization routine would have needed to find the starting time for each

activation of each tract variable, when it was turned off once it was activated, the target, spring constant and damping coefficient during each activation, and the weights of the articulators involved. Here, instead of the spring constant and damping coefficient, the equivalent natural frequency and damping ratio were specified. See Table 1 for a list of parameters necessary to specify each activation of a tract variable. There is a potentially infinite set of parameters that needs to be searched, given no limit to possible activations. The restrictions that were put on this potential were partly a result of the natural constraints provided by the genetic algorithm described below. In the present work, it was assumed that the articulators were recruited for each tract variable with known articulator weightings, so that the weightings were not recovered in the optimization. Some further constraints imposed to reduce the number of unknown parameters in this optimization will be described in Section 3.

In contrast to articulatory variables, the tract variables describe geometric features of the shape of the vocal tract tube in terms of constriction degree and location. Boë et al. (1992) have noted that acoustic output, in terms of formant frequencies, seems to be most sensitive to place and degree of constriction and emphasized that this fact should be used in acoustic-to-articulatory mapping. This is not surprising, as Mermelstein (1967) and Flanagan et al. (1980) have noted, the coefficients of the Webster horn equation, which is a good approximation to sound propagation in the vocal tract, are functions of the logarithm of the area function, thus making constriction area and position important determinants of resonance frequencies. Small changes in constriction degree and location can have a greater effect on resonance properties than small changes in the other parts of the area function. Early constrained area function models of the vocal tract (e.g. (Stevens and House, 1955)) recognized the importance of the constriction, and the articulatory model used by Meyer et al. (1989) is largely specified with constrictions. The results of Papçun et al. (1992), in their neural network computation, showed that the articulator involved with

the formation or breaking of a constriction was the most consistently tracked articulator. One of the reasons for optimizing for the transformation from acoustics to tract variables, rather than for the transformation from acoustics to articulators, in this work is that the tract variables specify the acoustically salient features of the area function more directly than do the articulators. While it may be true that there is strictly only one articulatory specification for a given area function, there may be many and disparate sets of articulatory coordinates that are close by, in the sense of producing similarly placed vocal tract constrictions. For instance, an alveolar constriction can be specified with varying amounts of jaw, tongue body and tongue tip displacement, because these articulators may compensate for each other to attain the prescribed constriction degree and location. While there is not complete compensation throughout the vocal tract, the constriction degree and location can be preserved using compensation. The importance of compensatory activity has been shown experimentally in human speech, e.g. (Abbs and Gracco, 1984; Kelso et al., 1984). Future developments involving the inclusion of aerodynamic sound source information in the articulatory synthesizer and into the acoustic parameter list for inversion may also make the importance of degree and place of constriction more apparent.

In the present case it was correct to assume known articulatory weights because the model used to produce the data was the same one used in the recovery procedure. This assumption would have to be scrapped when using this recovery method on actual data. To recover the articulator weightings and task-dynamic parameters may be much more difficult than recovering task-dynamic parameters alone. Not only are more parameters variable in the optimization, but the reason for recovering task-dynamic parameters in the first place was that constrictions are the most acoustically salient geometric feature. The contribution of each articulator to the formation or breaking of constrictions is, therefore, more sensitive to noise in the data. Because the method of recovery presented here can make use of speech physiology and production data, it may be possible to

ask subjects to perform relatively simple tasks, where the tract variables and targets are known, and to optimize first for articulatory weights. Such a procedure would be preliminary to more full-scale recovery.

Overall, the method used here was analysis-by-synthesis with an extra level: task dynamics of tract variables. Using geometric transformations, task dynamics was used to compute articulatory trajectories, which produced acoustic trajectories using ASY. In turn, the acoustic trajectories were compared with the formant data trajectories for improvement using a genetic algorithm for optimization. From the resulting optimum solution in terms of task-dynamic parameters, articulatory trajectories could be recovered simply by using the task-dynamic simulation.

2.2. *The genetic algorithm*

The relations between the task-dynamic parameters and the formant frequencies were much too complicated to write explicitly and to permit the finding of partial derivatives of the error function with respect to the task-dynamic parameters. Also, given the potentially high dimensionality of the space of task-dynamic parameters, many function evaluations would have had to be performed to estimate partial derivatives of the error measure. Function evaluations were very costly because they involved integrating a coupled set of nonlinear differential equations over the time interval of a /CV/ utterance. Therefore, methods involving derivatives, such as gradient descent (e.g. (Shirai and Kobayashi, 1986)), were avoided. The specific properties that made a genetic algorithm useful will be discussed after its description below.

The particular genetic algorithm employed for this study was a slightly modified version of an algorithm described by Goldberg (1989). In the algorithm used here, the individuals of a population were assigned randomly chosen task-dynamic parameter sets that were coded into binary strings called “chromosomes”, and each was assigned a fitness. The fitness used in this study was the inverse of the error measure already defined (see Eqs. (1) and (2)). Individuals were 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 mated 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 were evaluated based on their parameter sets as coded in their chromosomes. That is, the task-dynamic model was run based on the parameters specified by each child's chromosome, and the resulting fitness was computed according to Eq. (2) for each child. A small probability of mutation was allowed. As described so far, the algorithm was the Simple Genetic Algorithm given by Goldberg (1989, pp. 59–70). In a variation of this genetic algorithm, the best individual of a given generation was always retained into the succeeding generation.

It can be shown that patterns of individuals, called schemata, that are somewhat above average in fitness tend to increase in number from one generation to the next (Goldberg, 1989, pp. 28–33). There is a tendency for genetic algorithms to produce more individuals of higher fitness in succeeding generations. The power of the genetic algorithm comes from what John Holland, the originator of genetic algorithms, called *implicit parallelism*. Although a population of chromosomes is finite, say size N , the number of schemata being processed is on the order of N^3 (Goldberg, 1989, pp. 40–41). This implicit parallelism, as well as the probability of mutation, makes the algorithm much less likely to become stuck in local fitness maxima. Also, the implicit parallel processing property makes this algorithm more efficient than an exhaustive search done without the benefit of a codebook that provides reasonable starting values. Parallel processing in the usual sense of using many processors is also possible. Given that children have been produced as the result of mating a generation of individuals, the children's fitnesses can be evaluated on physically distinct processors. This capability was not used in the present work.

A genetic algorithm technique was chosen because it was readily implemented and its evolu-

tion during the optimization procedure was relatively easy to follow, allowing for tuning of the algorithm's specifications, such as population size and rate of mutation. A genetic algorithm requires that the parameters for optimization be coded into finite length strings (the chromosomes), which limits the range of any parameter and essentially discretizes the parameter space. The degree of discretization of each parameter can be varied to tune the optimization. This was controlled by varying the range of allowed parameters and the number of bits given to code a specific parameter. Because the ranges of starting and ending activation times were both limited to discrete steps and finite range, the potentially infinite set of parameters was made finite. Further, the ease with which one could limit the ranges of other parameters to constrain the optimization was a very attractive feature because of the difficulties in running the task-dynamic model beyond certain limits (e.g. high natural frequencies). Genetic algorithms for optimization might be classified as stochastic, and, as noted above, there is some precedence for stochastic techniques in the generation of codebooks of articulatory-acoustic correspondences by random sampling of articulation (Schroeter et al., 1990). A major difference, of course, is that the procedure described here is performed on-line as opposed to off-line, as in the case of the codebook technique. A non-stochastic procedure, such as that used by Parthasarthy and Coker (1992) was a possible alternative, and it remains to determine which optimization procedure is the best.

The coding used here was a simple binary code for the real number, task-dynamic parameters. The coded parameters were concatenated to form a complete chromosome. A better way of forming the chromosomes might have been to split the binary representations of the parameters so that the most significant bits of all the parameters were grouped together, then the next significant bits are grouped together, and so on. This may have been a better method because of the greater likelihood that the shorter lengths of chromosomes would stay together through the mating process and that the fitness of any individual depends on how the parameters interact.

3. Procedure

A program, ASYINV, was written to achieve articulatory recovery using the method described in the previous section. It was necessary to simulate the formant trajectory data. This was accomplished by means of a gestural score (Browman and Goldstein, 1990); a specification of task-dynamic parameters, including the activation times, natural frequency, damping ratio, target position, and the weights of the various articulators involved for each of the tract variable activations that compose the utterance. A gestural score could be written into a computer file to be used as input to the task-dynamic simulation. This, in turn, could generate the tables that enabled the articulatory synthesizer (ASY) to estimate the formant frequency values.

Constraining relations were used to keep the number of unknown parameters in B , K and z_0 , as well as activation times, to a minimum. It was assumed that all movements were critically damped, and that the activation intervals were equal to the period computed from natural frequency. Also, the activation intervals were assumed to be at least 100 ms long to avoid movements that were too stiff for the task-dynamic simulation. With these constraints, the unknowns for a given tract-variable activation were the beginning and ending activation times and the target position (see Table 1). There could have been more than one activation of any tract variable, and these could have overlapped in time, although there was never more than one activation interval for a tract variable with unknown parameters in these tests. Also, the actions of the tract variables were grouped so that some would have identical activation intervals. The tongue body tract variables TBCL and TBCD were in one such group, TTCL and TTCD were in another group, and LP and LA were in a third group (see Table 1).

Gestural scores were designed to produce articulatory movement for utterances that resembled /əbæ/ and /ədæ/. The gestural scores were generated from the linguistic gestural model of Browman and Goldstein (1990), and then modified to meet the constraints noted in the previ-

ous paragraph. The trajectories of the tract variables, activation intervals, and target positions specified by the scores for /əbæ/ and /ədæ/ are illustrated in Figs. 3 and 4. In the recovery process, some parts of the gestural score were taken as known and fixed. In the case of /əbæ/,

for the initial movement to lip closure involving the tract variables LA and LP, the activation interval and targets were taken as known (see Table 2). However, the subsequent activation interval for LA and LP was taken as unknown, as was the target position for LA. Only the target

Table 1
Task-dynamic parameters to specify each activation

Tract variable	Variable task-dynamic parameters without constraints		Variable task-dynamic parameters with constraints	
LA (lip aperture)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position	start activation time end activation time (> 100ms + start time) (natural frequency and damping are derived from start and end activation time)
LP (lip protrusion)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position	
TBCL (tongue body constriction location)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position	start activation time end activation time (> 100ms + start time) (natural frequency and damping are derived from start and end activation time)
TBCD (tongue body constriction degree)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position	
TTCL (tongue tip constriction location)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position	start activation time end activation time (> 100ms + start time) (natural frequency and damping are derived from start and end activation time)
TTCD (tongue tip constriction degree)	start activation time end activation time natural frequency	damping ratio target position articulator weights	target position	

Table 2
Fixed and variable task-dynamic parameters for the recovery of /əbæ/

Activated tract variable	Task-dynamic parameters	
LA } first activation	<u>target position</u>	start activation time
LP } first activation	target position	end activation time
LA } second activation	<u>target position</u>	start activation time
LP } second activation	target position	end activation time
TBCL	<u>target position</u>	start activation time
TBCD	<u>target position</u>	end activation time
	fixed parameters	variable parameters

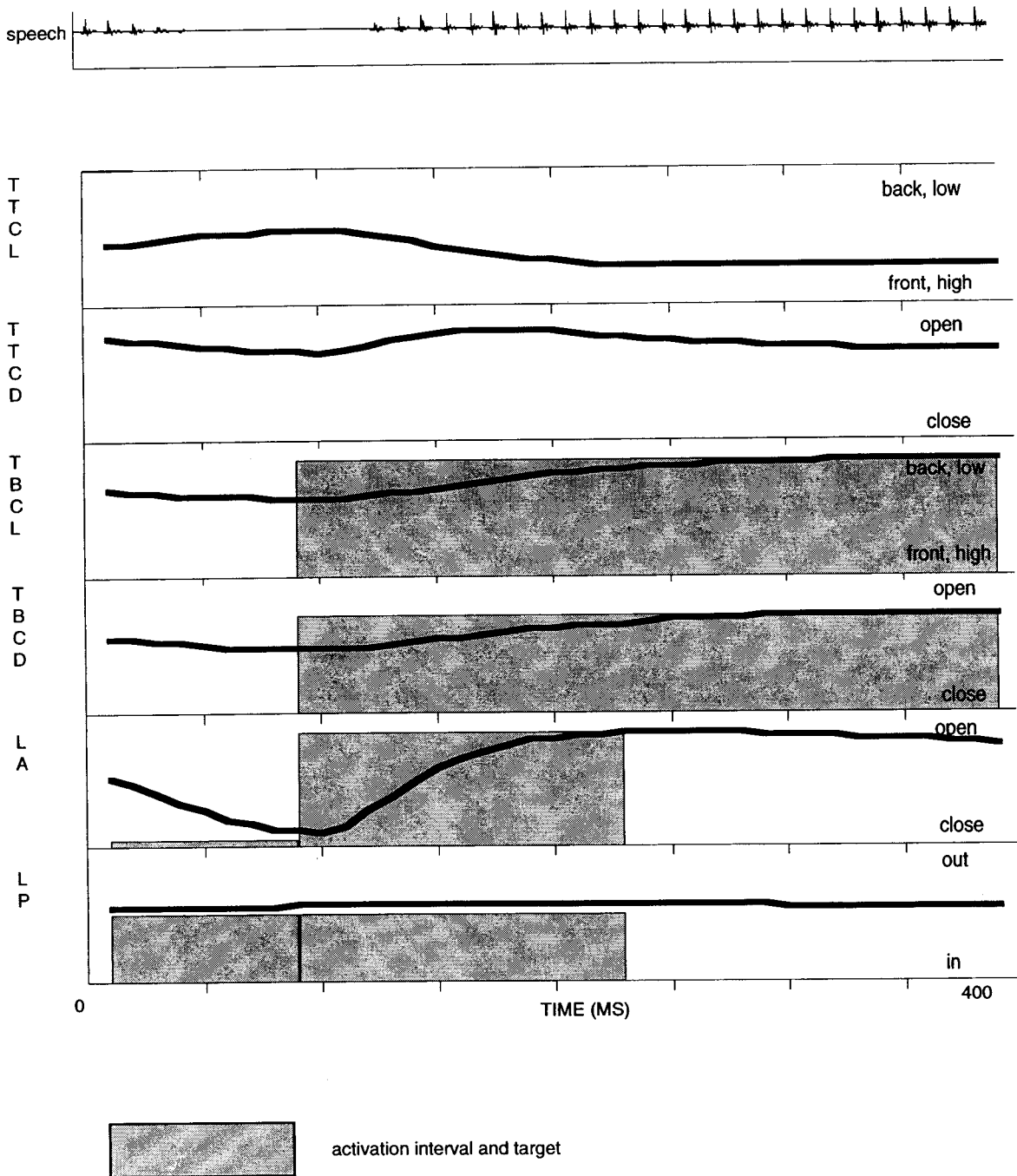


Fig. 3. Tract variable trajectories, activation times and targets for utterance /əbæ/. The activation times are shown by the length of the shaded boxes and the targets by the heights. The maximum possible value for TBCL is 3.49 radians. The maximum possible value for TTCL is 1.22 radians. The minimum 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.

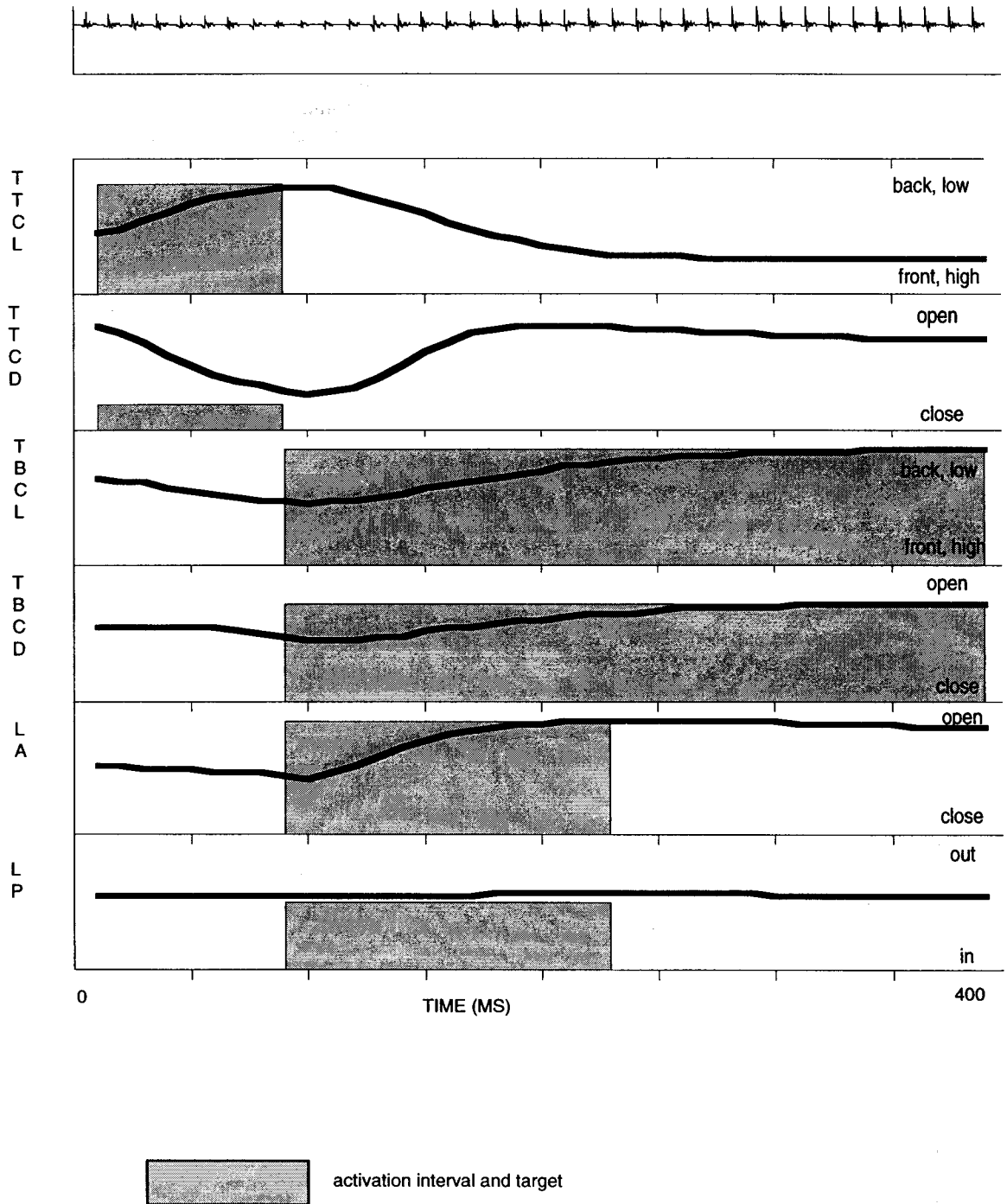


Fig. 4. Tract variable trajectories, activation times and targets for utterance /əɒæ/. The activation times are shown by the length of the shaded boxes and the targets by the heights. Ordinate scales as in Fig. 3.

position for LP was assumed known in its second activation. (It was not possible to recover LP and LA simultaneously with information only from the first three formants.) Tongue body movement was taken as unknown, so that the activation interval and the target positions for TBCL and TBCD were varied for the optimum fit. The fact that TTCL and TTCD were not activated for /əbæ/ was also assumed to be known. For /ədæ/, the activation times and targets of the tongue-tip tract variables, TTCL and TTCD, for the alveolar closure were presumed unknown (see Table 3). The dynamics of LA and LP for the final vowel, which occurs after the tongue-tip closure, were taken as completely known. The parameters of the tongue body tract variables in the transition to final vowel, TBCL and TBCD, were taken as unknown in this instance. Thus, the tests on /əbæ/ differed from those on /ədæ/ in that unknown activations were allowed in sequence in the latter case, but not in the former.

Given the synthetic acoustic data, the program ASYINV was used for articulatory recovery un-

der two conditions; one where no noise was added to the formant frequency data, and another where random noise with a flat distribution between -10 and $+10$ Hz was added to each formant frequency at each time frame. For each test an initial population of chromosomes for 60 individuals was generated using a random number generator. As described in Section 2, their fitnesses were evaluated by using the task-dynamics parameters specified by each individual to produce a kinematic description for driving the articulatory synthesizer, ASY, which, in turn, created formant trajectories. The fitness of an individual was the inverse of the sum of squares of the differences of each of the lowest three formant frequencies produced by the individual and that of the data in 10 ms steps (see Eqs. (1) and (2)). Table 4 indicates the range of the target values for each tract variable, the number of bits used in the coding of that parameter into the chromosomes, and the resulting resolution of that target. Beginning and ending times for the activation intervals were resolved within the data frame rate

Table 3
Fixed and variable task-dynamic parameters for the recovery of /ədæ/

Activated tract variable	Task-dynamic parameters	
LA	target position	start activation time
LP	target position	end activation time
TBCL	target position	start activation time
TBCD	target position	end activation time
TTCL	target position	start activation time
TTCD	target position	end activation time
	fixed parameters	variable parameters

Table 4
Target value specifications

Tract variable	Maximum/minimum target value	Number of bits in chromosome	Resolution
LA	1.80/–0.3 cm	6	0.033 cm
TBCL	3.16/0.51 rad	6	0.042 rad
TBCD	{ /əbæ/ { /ədæ/	{ 1.80/–0.30 cm { 1.63/–0.13 cm	{ 0.033 cm { 0.028 cm
TTCL	{ without noise { with noise	{ 1.27/0.31 rad { 1.16/0.40 rad	{ 0.015 rad { 0.012 rad
TTCD	2.15/–0.65 cm	6	0.044 cm

of 10 ms. Thirty pairs of individuals were chosen with probability proportional to each of their fitnesses. There was a 0.6 chance of each of these pairs to mate. There was also a 0.001 chance of mutation to a single position in any of the children's chromosomes. If mating did occur, the fitnesses of the children strings would then be evaluated. The choice of pairs and possible mating was allowed to continue for 60 generations. At the end of 60 generations the individual with the greatest fitness had its string decoded into task-dynamic parameters, which were stored for later comparison. This procedure was repeated 8 times, and the fittest individual of all the runs was taken as the best approximation to the task-dynamic parameters. These parameters could be

used to drive ASY and the articulator trajectories of the best individual could be compared to those that generated the original data.

4. Results and discussion

The results are shown here in three domains: the articulatory, the tract variable and the acoustic. While the purpose of these experiments on a model was to find how well the proposed method could recover articulation under ideal conditions, the other domains can be studied to help evaluate the results. A sampling of the results is shown in Figs. 5–17, for both no-noise and noise conditions.

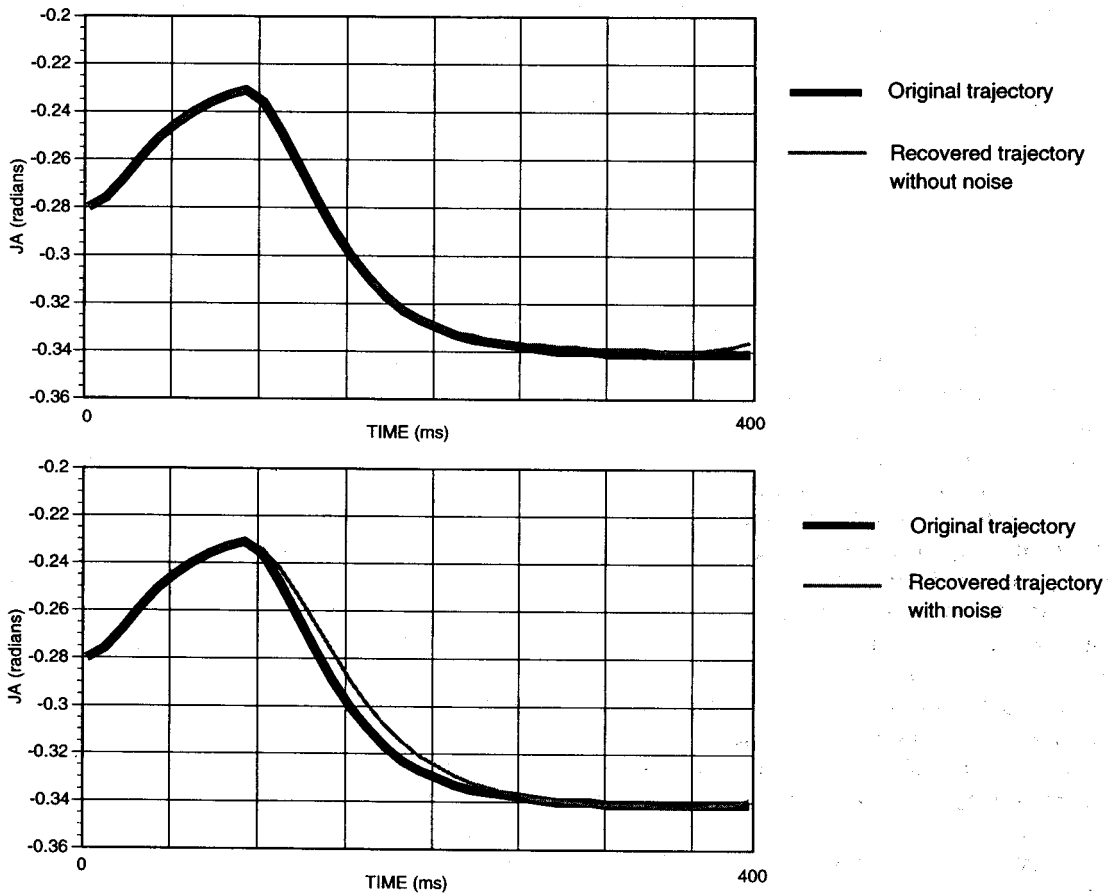


Fig. 5. Original and recovered jaw vector angle (JA) trajectories for utterance /əbæ/.

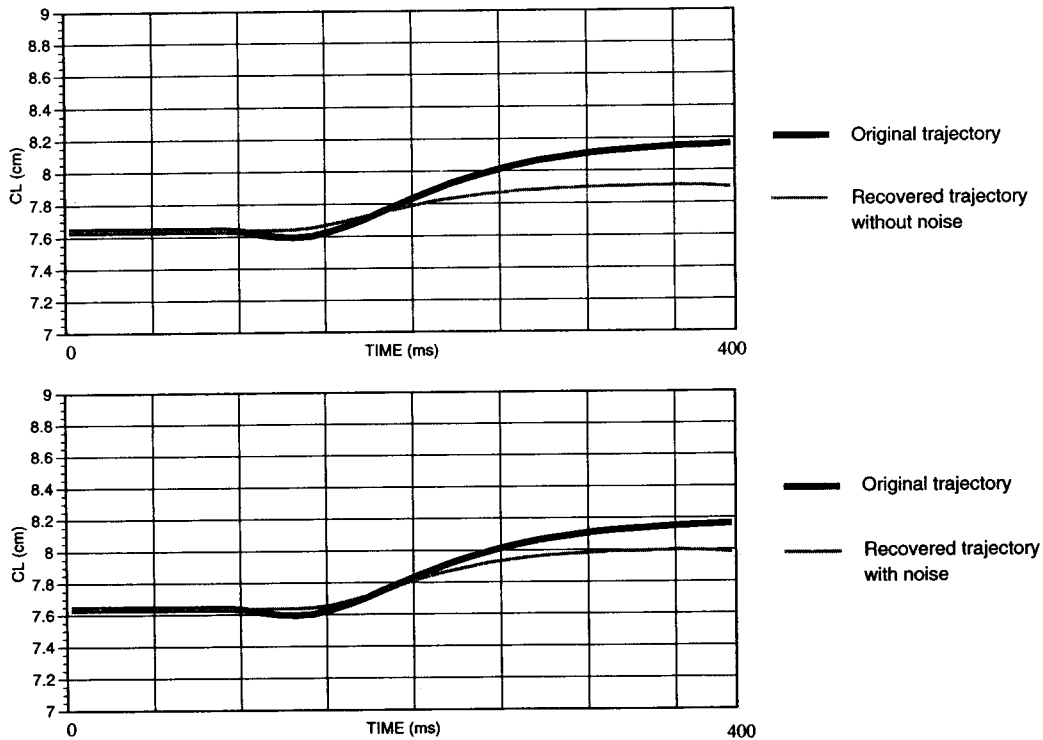


Fig. 6. Original and recovered tongue body center vector length (CL) trajectories for utterance /əbæ/.

The recovered and original trajectories for the utterance /əbæ/ (Figs. 5–10) were identical for the first 90 ms, or so, because the task dynamics were presumed known for that interval. Also, it can be seen in these figures that the recovery using noisy data was as good as that for perfect data. The mean (over time) square error in the sum of the differences of the recovered and original formant frequencies was 1120 for recovery without noise and 1320 for recovery in noise. This corresponds to an average error of about 19 Hz for each formant frequency at each time frame in the no-noise condition and of about 21 Hz in the noise condition. Note that the standard deviation of the noise distribution was 5.8 Hz.

In the tests on /ədæ/ there was no part of the utterance where the task-dynamic parameters were presumed to be completely known, which explains why the recovered trajectories did not match the originals perfectly from the beginning. Figs. 11–17 show that the recovery was not af-

fectured much by the noise added to the formant frequency data. The average error per formant frequency data point was about 20 Hz for recovery without noise and it was 28 Hz for recovery in noise.

A more complete, quantitative description of the results is given for the utterances /əbæ/ and /ədæ/ in Tables 5 and 6, with the root mean square and maximum absolute errors in the articulator trajectories. All maximum errors were less than 2 mm for length measures and less than 0.1 radians for angle measures. A detailed description of the results shown in Figs. 5–17 will complete the discussion.

Jaw angle for /əbæ/ was recovered almost perfectly in both the no-noise and noise condition (Fig. 5). The errors made in the recovery of tongue body center vector length, CL, (Fig. 6) and tongue body center vector angle, CA, (Fig. 7) for the utterance /əbæ/ were due to errors made in recovering the task-dynamic parameters.

This is illustrated by considering the recovery in the no-noise condition as an example. A review of the recovered gestural score shows that the recovered tongue body constriction location (TBCL) target was slightly higher and more forward than the target of TBCL in the original data producing gesture, and that the recovered tongue body constriction degree (TBCD) was more closed in the recovered than in the original gesture (Fig. 8). Further, the activation intervals for these tract variables started later and were of shorter duration than the original. Thus, there was less time to move the tongue body constriction into the pharynx and to open the tongue body constriction after the lip opening in the recovered utterance. Given that jaw angle, JA, was recovered almost

perfectly, all of these facts would indicate a shorter tongue body center vector length, CL, and smaller tongue body center vector angle, CA, as the final vowel gesture progressed than in the original utterance, because TBCL and TBCD depend on the articulators JA, CL and CA (see Fig. 2). Figs. 8 and 9 indicate that the task-dynamics for the lip aperture was not recovered as well in the noise condition as it was in the no-noise condition. There is no appreciable difference in how well the recovered formant trajectories fit the original in the no-noise and noise conditions (Fig. 10).

As the numerical testing progressed, it became clear that simultaneously attempting to obtain lip protrusion (LP) target and lip aperture (LA) tar-

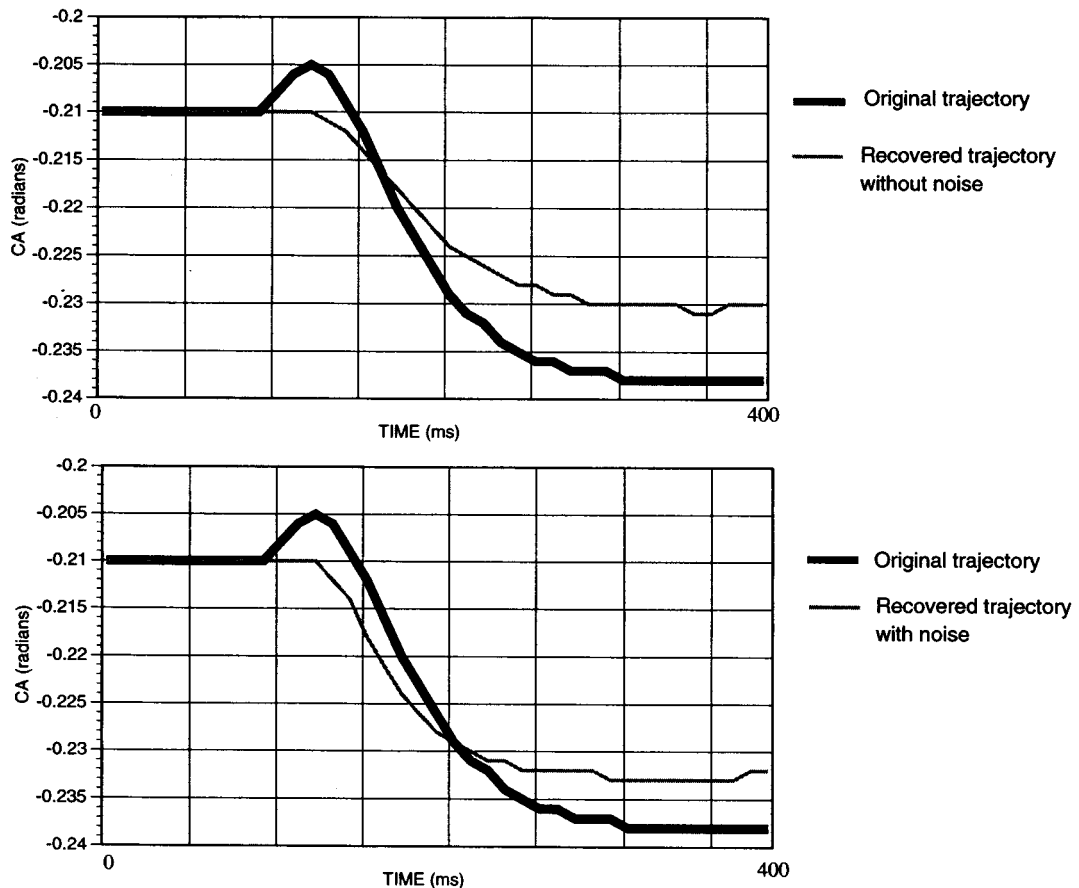


Fig. 7. Original and recovered tongue body center vector angle (CA) trajectories for utterance /əbæ/.

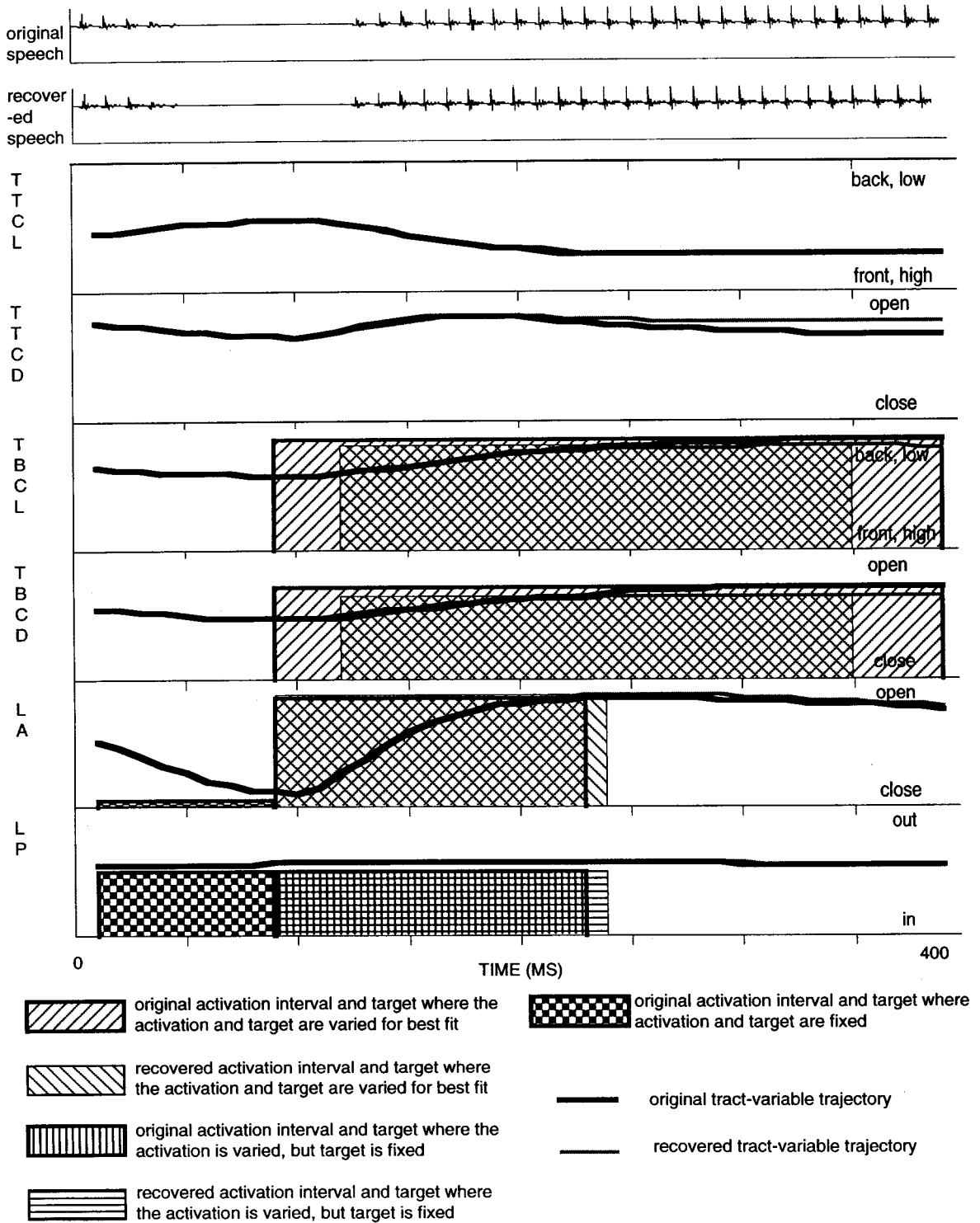


Fig. 8. Original and recovered tract variable trajectories, activation times, and targets for utterance /əbæ/ without noise in the data. The activation times are shown by the length of the shaded boxes and the targets by the heights. Ordinate scales as in Fig. 3.

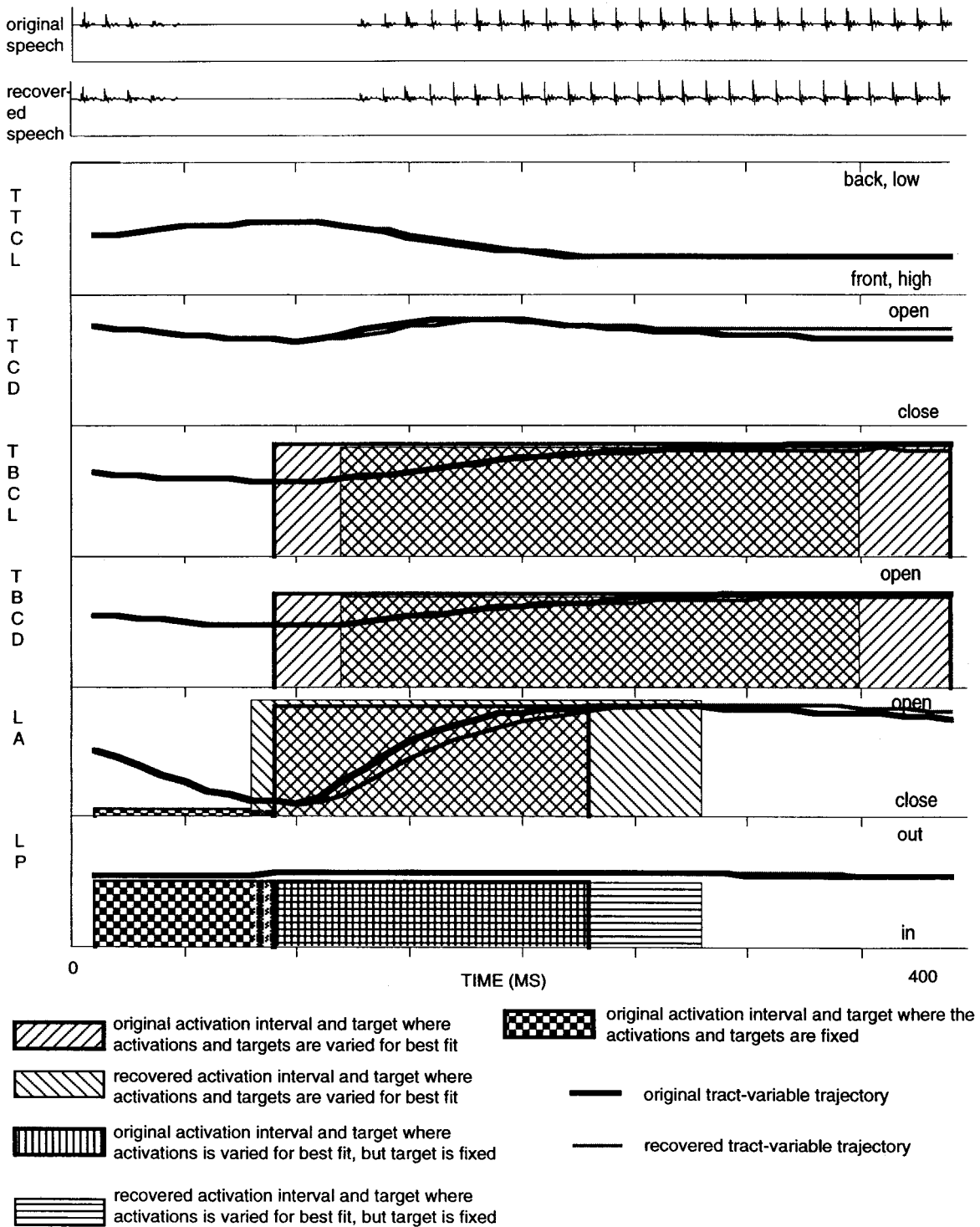


Fig. 9. Original and recovered tract variable trajectories, activation times, and targets for utterance /əbæ/ with noise in the data. The activation times are shown by the length of the shaded boxes and the targets by the heights. Ordinate scales as in Fig. 3.

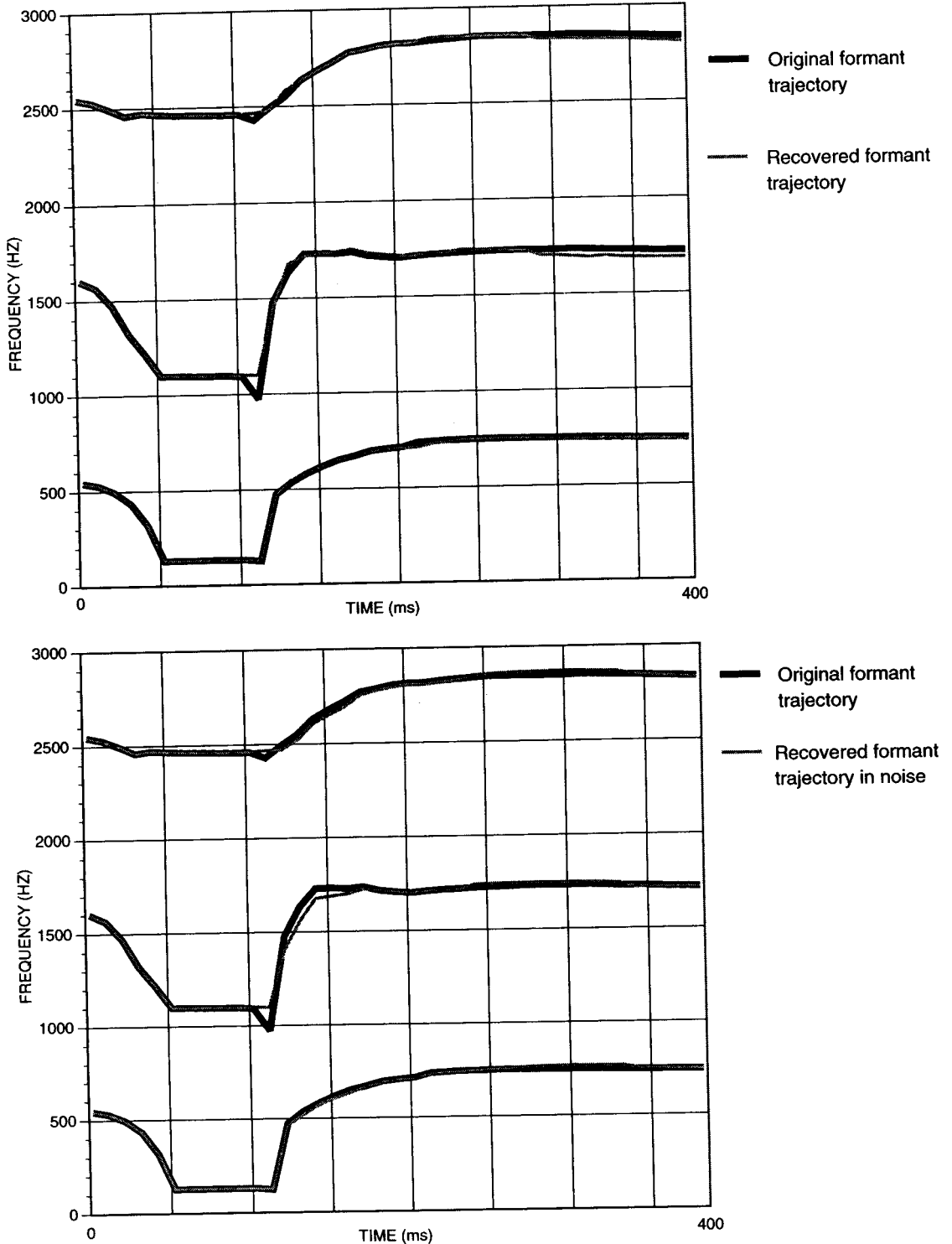


Fig. 10. Original and recovered first three formant trajectories for the utterance /əbæ/.

get was too difficult given three formant frequency trajectories, because protruding the lips was equivalent to decreasing lip aperture given this limited set of acoustic parameters. However, even after the protrusion target was specified, there was still trouble in obtaining LLV and ULV articulations. The reason for this can be seen in the utterance / $\text{əd}\text{æ}$ /. Here the task-dynamics of the lip aperture, LA, was assumed known, so that any error in the jaw angle (JA), perhaps due to errors in the recovery of the task dynamics of another tract variable, should have caused errors in the upper and lower lip vertical positions (ULV and LLV) positions, because these are the three articulators involved with LA (Fig. 2). Because the jaw in the recovered movement did not open enough (Fig. 11), the compensatory activity of the

LLV can be seen in Fig. 12. The lower lip opens further in the recovered trajectory than in the original movement to make up for the lack of jaw movement.

The gestural score recovered for / $\text{əd}\text{æ}$ / for the no-noise condition is shown in Fig. 15 and for the noise condition in Fig. 16. In both conditions, it is apparent that the target tongue tip constriction degree, TTCD, was not as tight as the original, with the recovery in noise furthest from the original. The consequences of this on the tongue tip angle articulator, TA, can be seen in Fig. 13, where the maximum angle in the closure region was underestimated, again with the recovery in noise being furthest from the original. The underestimate in TTCD target may have allowed the tongue body activation to be delayed relative to

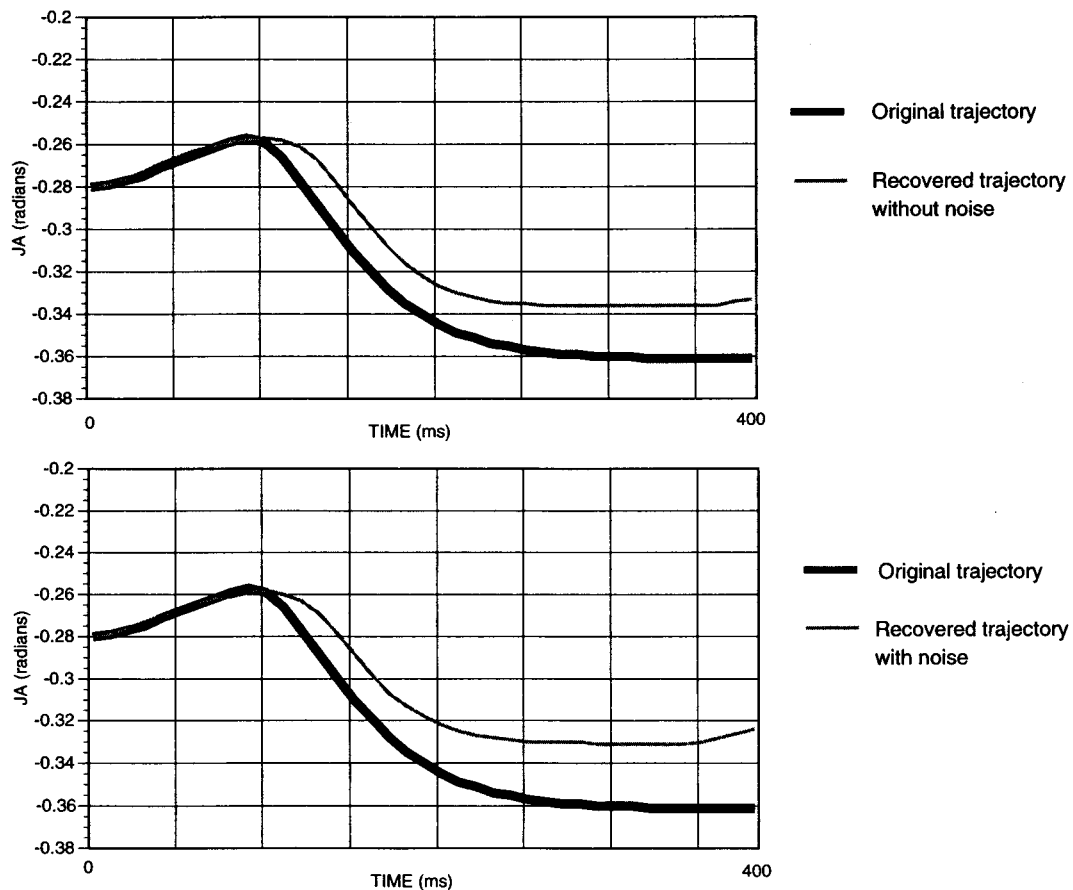


Fig. 11. Original and recovered jaw vector angle trajectories (JA) for utterance / $\text{əd}\text{æ}$ /.

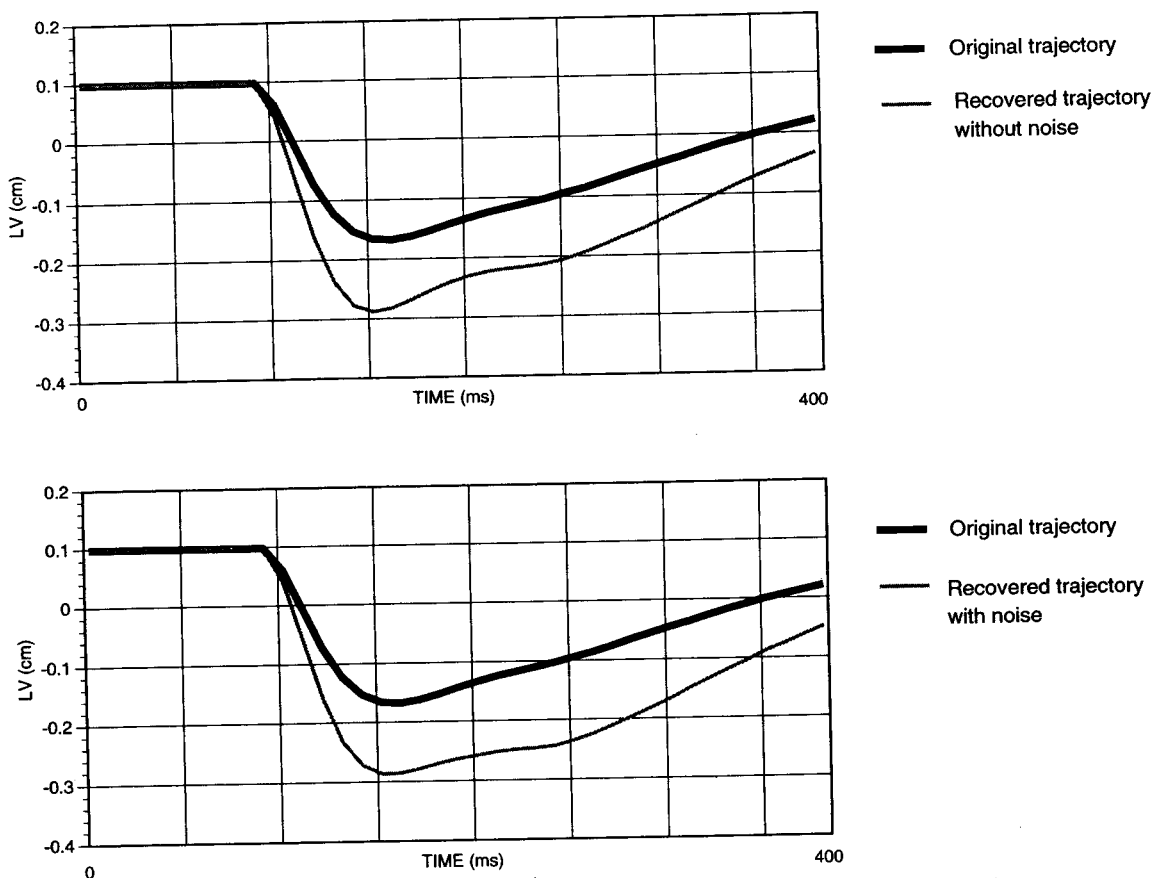


Fig. 12. Original and recovered lower lip vertical position (LLV) trajectories for utterance /ædæ/.

the tongue-tip release, because the underestimate in TTCD target meant that the formant trajectories did not have as far to go to reach the steady state in the second vowel. This delay can be seen at the articulatory level in the recovered jaw angle, JA, (Fig. 11), recovered tongue body center vector length, CL, (Fig. 14) and the recovered third formant (Fig. 17). Compensation at the tract variable level, where a spatial error in one gesture is coupled with a temporal error in another gesture, may be alleviated using more acoustic information. In particular, inclusion of aerodynamic sound sources may alleviate ambiguities in the degree and timing of gestures involving tight degrees of constriction because of the sensitivities of aerodynamic sources to small variations in constriction degree. This information would allow

the algorithm to determine the time when a tight constriction has been released. Timing being more precisely determined may force a better determination of the target.

5. Conclusion

The results indicate that the proposed method of articulatory recovery is worth pursuing with further model testing and also testing using real speech. There has been some success in obtaining articulatory trajectories with just the first three formant frequency trajectories, but it is apparent that other acoustic information must necessarily be added for later use, for instance to obtain lip aperture and lip protrusion simultaneously. There

has been no particular one-to-many problem here, and it can be tentatively concluded that this method is highly enough constrained to avoid such problems, as with previous work, see (Schroeter and Sondhi, 1992; Parthasarthy and Coker, 1992). The constraints in the present work were provided by a model vocal-tract and task-dynamics. Task-dynamics has the further advantage of controlling the acoustically salient geometric features of constriction degree and location. A genetic algorithm was relatively easy to implement as an optimization procedure. Genetic algorithms are stochastic procedures that do not require derivatives, and they provide a natural means of performing a bounded optimization through the coding of parameters.

As can be expected, there will be complications in moving from model tests to actual human speech. One problem is that of removing the assumption of known weightings of the articula-

tors used by a tract variable. Acoustic data will have to be derived from a real speech wave, instead of from a computed transfer function. Further, such constraints as critical damping and activation intervals equal to the period based on natural frequency may also be removed. Further, the model vocal tract will need to be customized in terms of length and shape for each subject for later recovery. These are large problems, but a step-by-step approach will lead to the best possible solution. In the following, proposed solutions to problems encountered in the model tests and methods for moving to human speech will be outlined.

In the model tests, there were particular problems in recovering the relative timing of activation intervals in the task dynamics. However, using acoustical information other than the three formant frequency trajectories may help in recovering the critical timing information, because re-

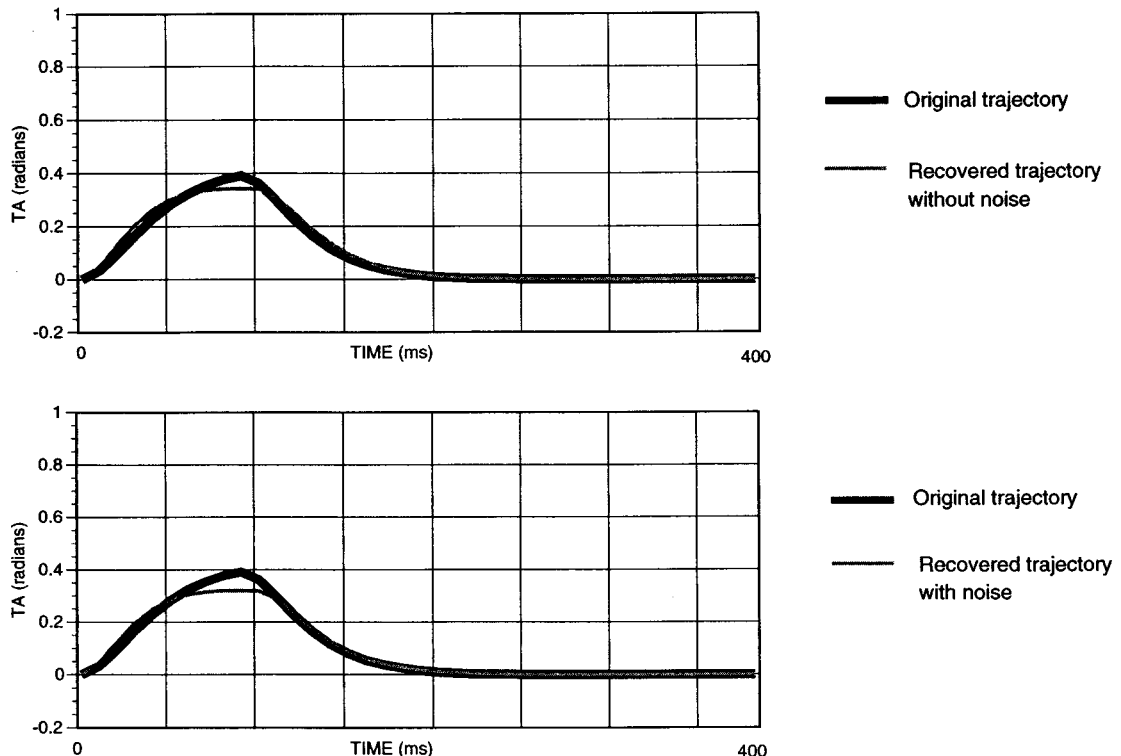


Fig. 13. Original and recovered tongue tip vector angle (TA) trajectories for utterance /ədæ/.

leases and closures are often marked by abrupt changes in source amplitude and type. In other words, source information may be a good addition to the formant frequency information because it may be the least redundant addition to formant information for the purposes of articulatory recovery. This will require that aerodynamic quantities be included into task dynamics, and that a pressure wave be synthesized so that such quantities as RMS amplitude can be computed. This will also mean that the fitness function will incorporate information of a different physical dimension than frequency, or inverse time.

As far as modifications for human subjects are concerned, initially, it is proposed to use a customized model vocal tract. This vocal tract model would be molded to the dimensions of an individual based on data derived from the various static

means of measurement. An example of a direct measure would be one obtained from an MRI scan (Baer et al., 1991), and an indirect method would be one in which the impedance tube technique of Sondhi and Resnick (1983) or of Milenkovic (1987) is used. An unknown in the current research is how to process the acoustic data for robust estimation, although the findings of others may be of help here (Shirai and Kobayashi, 1986; Meyer et al., 1991). Perhaps the most difficult problems will be encountered when obtaining the weightings of the articulators for the tract variables and modifications of task-dynamics for fast movements near obstruents (Stevens, 1993). To obtain the weightings, some simple movements of the tongue or lips may be required, so that only selected tract variables are activated. The issue of the adequacy of task-dy-

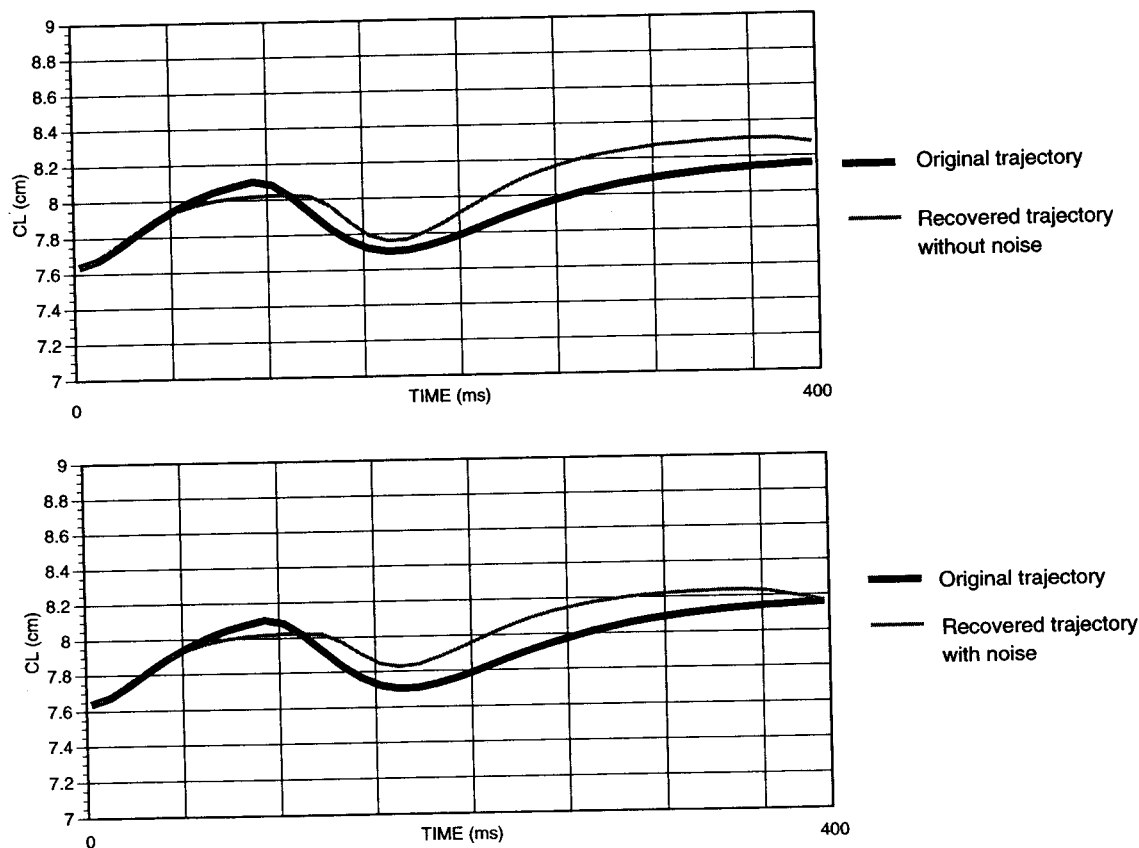


Fig. 14. Original and recovered tongue body vector length (TL) trajectories for utterance /ædæ/.

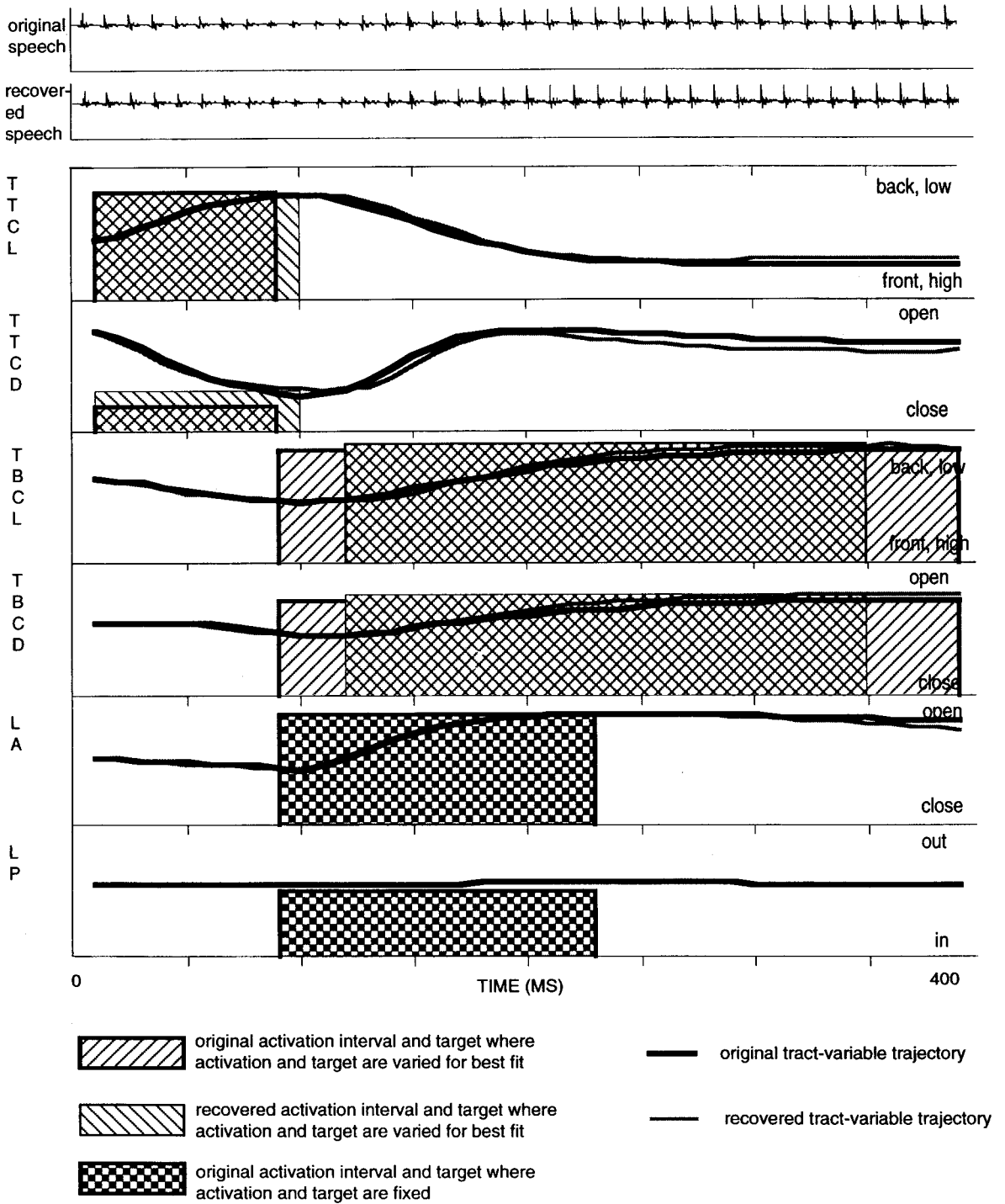
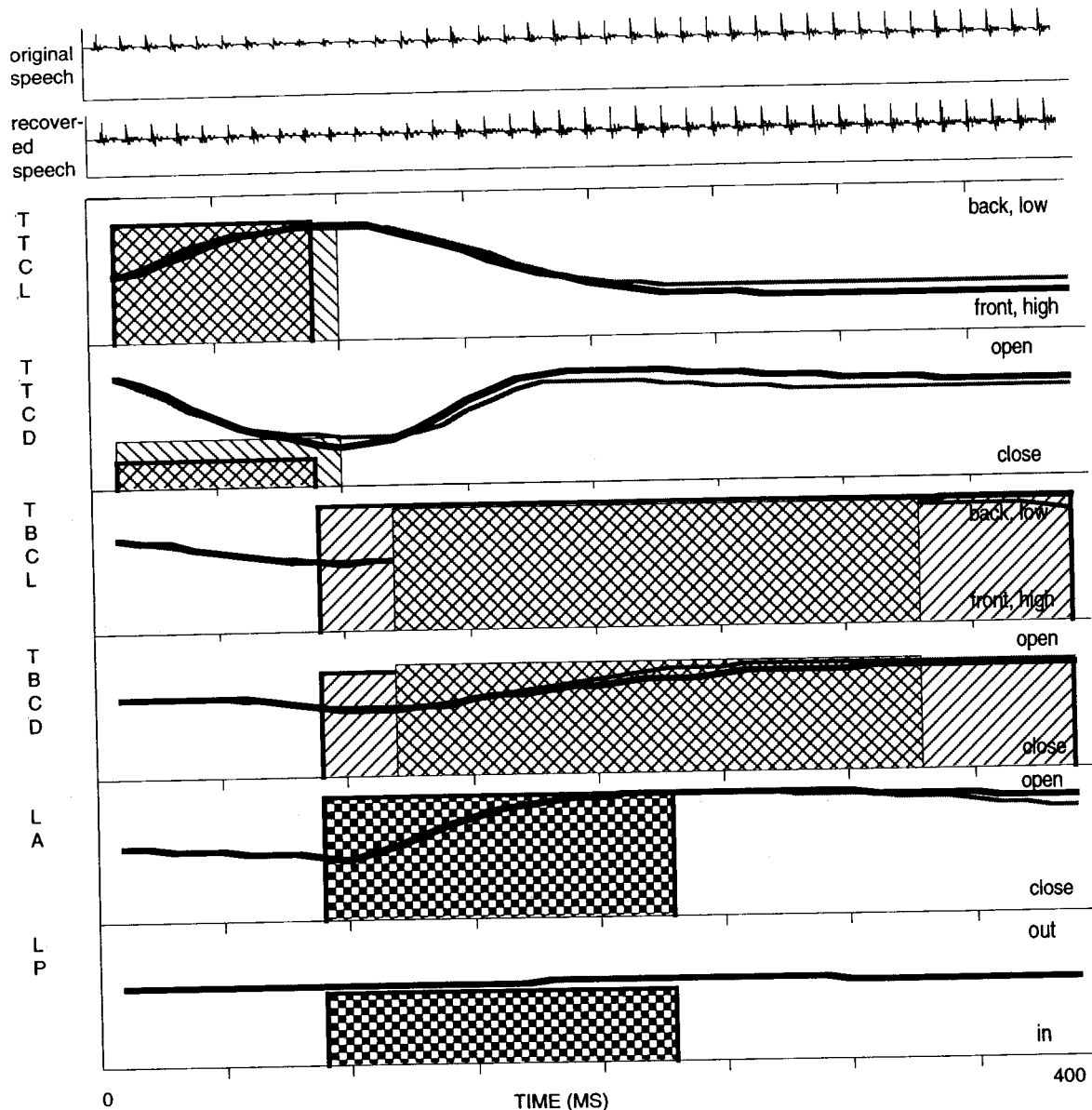





Fig. 15. Original and recovered tract variable trajectories, activation times and targets for utterance /ədæ/ without noise in the data. The activation times are shown by the length of the shaded boxes and the targets by the heights. Ordinate scales as in Fig. 3.



 activation interval and target where the activation and targets are fixed

 original activation interval and target where the activation and targets are varied for best fit

 recovered activation interval and target where the activation and targets are varied for best fit

 original tract-variable trajectory


 recovered tract-variable trajectory

Fig. 16. Original and recovered tract variable trajectories, activation times, and targets for utterance /ædæ/ with noise in the data. The activation times are shown by the length of the shaded boxes and the targets by the heights. Ordinate scales as in Fig. 3.

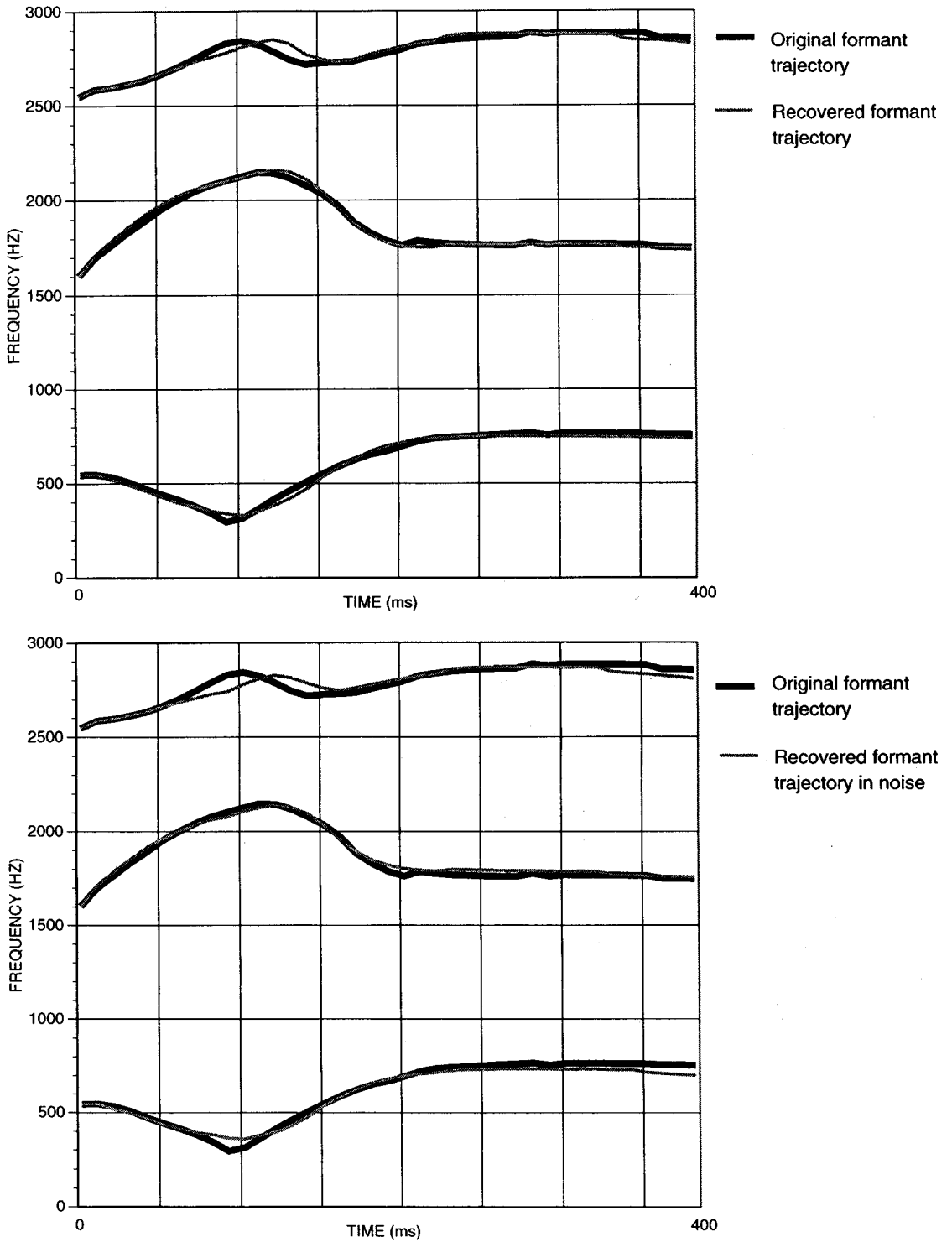


Fig. 17. Original and recovered first three formant trajectories for the utterance /ədæ/.

Table 5
Comparison of original and recovered articulator trajectories for utterance /əbæ/

Articulator coordinate	RMS difference (without noise)	RMS difference (with noise)	Maximum absolute difference (without noise)	Maximum absolute difference (with noise)
CL	0.021 cm	0.013 cm	0.27 cm	0.18 cm
CA	0.00086 rad	0.00061 rad	0.0080 rad	0.0060 rad
JA	0.00024 rad	0.00075 rad	0.0050 rad	0.012 rad
LP	0.00019 cm	0.00088 cm	0.0027 cm	0.0012 cm
LLV	0.0064 cm	0.014 cm	0.064 cm	0.16 cm
ULV	0.0063 cm	0.014 cm	0.064 cm	0.16 cm

Table 6
Comparison of original and recovered articulator trajectories for utterance /ədæ/

Articulator coordinate	RMS difference (without noise)	RMS difference (with noise)	Maximum absolute difference (without noise)	Maximum absolute difference (with noise)
CL	0.020 cm	0.018 cm	0.19 cm	0.18 cm
CA	0.0034 rad	0.0027 rad	0.031 rad	0.025 rad
TL	0.031 cm	0.0081 cm	0.18 cm	0.12 cm
TA	0.0027 rad	0.0031 rad	0.049 rad	0.072 rad
JA	0.0030 rad	0.0036 rad	0.028 rad	0.037 rad
LLV	0.013 cm	0.016 cm	0.13 cm	0.14 cm
ULV	0.13 cm	0.016 cm	0.13 cm	0.14 cm

namics to describe the movement of articulators over an entire obstruent-sonorant interval remains a research question.

Finally, the connection between this method and off-line techniques may be made a little closer. To save computation time it would make sense to convert the present technique to one that functions off-line such as the codebook or neural net techniques, e.g. (Schroeter et al., 1987; Paçun et al., 1992). Within the framework of genetic algorithms, classifier systems present themselves as a means of doing this (Goldberg, 1989). It remains a research question as to whether using a classifier framework is the best means to convert the procedure presented here to an off-line technique.

6. Acknowledgments

The author thanks Ronald Coifman, Louis Goldstein, Patrick Nye, Philip Rubin, Juergen Schroeter and an anonymous reviewer for their comments and discussions. The support of the

NIH through grant DC-01247 to Haskins Laboratories is also gratefully acknowledged.

7. References

- J.H. Abbs and V.L. Gracco (1984), "Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech", *J. Neurophysiology*, Vol. 51, pp. 705–723.
- B.S. Atal and S.L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer.*, Vol. 50, pp. 637–655.
- B. Atal, J.J. Chang, M.V. Mathews and J.W. Tukey (1978), "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *J. Acoust. Soc. Amer.*, Vol. 63, pp. 1535–1555.
- T. Baer, J.C. Gore, L.C. Gracco and P.W. Nye (1991), "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels", *J. Acoust. Soc. Amer.*, Vol. 90, pp. 799–828.
- L.-J. Boë, P. Perrier and G. Bailly (1992), "The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion", *J. Phonetics*, Vol. 20, pp. 27–38.
- C.P. Browman and L. Goldstein (1990), "Gestural specification using dynamically-defined articulatory structures", *J. Phonetics*, Vol. 18, pp. 299–320.

- C.H. Coker (1976), "A model of articulatory dynamics and control", *Proc. IEEE*, Vol. 64, pp. 452–460.
- H.K. Dunn (1950), "The calculation of vowel resonances, and an electrical vocal tract", *J. Acoust. Soc. Amer.*, Vol. 22, pp. 740–753.
- J.L. Flanagan, K. Ishizaka and K.L. Shipley (1980), "Signal models for low bit-rate coding of speech", *J. Acoust. Soc. Amer.*, Vol. 68, pp. 780–791.
- D.E. Goldberg (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA).
- J.L. Kelly and C.C. Lochbaum (1962), "Speech synthesis", in *Proc. 4th Internat. Congress of Acoustics*, Paper G-42, pp. 1–4.
- J.A.S. Kelso, B. Tuller, E. Vatikiotis-Bateson and C.A. Fowler (1984), "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures", *J. Experimental Psychology: Human Perception and Performance*, Vol. 10, pp. 812–832.
- J.N. Larar, J. Schroeter and M.M. Sondhi (1988), "Vector quantization of the articulatory space", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, pp. 1812–1818.
- S.E. Levinson and C.E. Schmidt (1983), "Adaptive computation of articulatory parameters from the speech signal", *J. Acoust. Soc. Amer.*, Vol. 74, pp. 1145–1154.
- A.M. Liberman and I.G. Mattingly (1985), "The motor theory of speech perception revisited", *Cognition*, Vol. 21, pp. 1–36.
- J. Liljencrants (1985), Speech synthesis with a reflection-type line analog, Doctoral dissertation, KTH, Stockholm, Sweden, unpublished.
- Q. Lin (1990), Speech production theory and articulatory speech synthesis, Doctoral dissertation, KTH, Stockholm, Sweden, unpublished.
- S. Maeda (1982), "A digital simulation method of the vocal-tract system", *Speech Communication*, Vol. 1, Nos. 3,4, pp. 199–229.
- R.S. McGowan (1991). "Recovering tube kinematics using time-varying acoustic information", *Proc. 12th Internat. Congress of Phonetic Sciences, Aix-en-Provence, 19–24 August 1991*, Vol. 4 (Aix-en-Provence, Université de Provence, Service des Publications), pp. 486–489; also in: Haskins Laboratories Status Report on Speech Research, SR-107–108, July–December, 1991, pp. 81–86.
- P. Mermelstein (1967), "Determination of the vocal-tract shape from measured formant frequencies", *J. Acoust. Soc. Amer.*, Vol. 41, pp. 1283–1294.
- P. Mermelstein (1973), "Articulatory model for the study of speech production", *J. Acoust. Soc. Amer.*, Vol. 53, pp. 1070–1082.
- P. Meyer, R. Wilhelms and H.W. Strube (1989), "A quasi-articulatory speech synthesizer for German language running in real time", *J. Acoust. Soc. Amer.*, Vol. 86, pp. 523–539.
- P. Meyer, J. Schroeter and M.M. Sondhi (1991), "Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks", *IEEE Trans. Signal Process.*, Vol. SP-39, pp. 1493–1502.
- P. Milenkovic (1984), "Vocal tract area functions from two-point acoustic measurements with formant frequency constraints", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 32, pp. 1122–1135.
- P. Milenkovic (1987), "Acoustic tube reconstruction from noncausal excitation", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35, pp. 1089–1100.
- G. Papçun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zacks and S. Levy (1992), "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data", *J. Acoust. Soc. Amer.*, Vol. 92, pp. 688–700.
- S. Parthasarthy and C.H. Coker (1992), "On automatic estimation of articulatory parameters in a text-to-speech system", *Computer Speech and Language*, Vol. 6, pp. 37–75.
- J.S. Perkell, M.H. Cohen, M.A. Svirsky, M.L. Matthies, I. Garabieta and M.T.T. Jackson (1992), "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", *J. Acoust. Soc. Amer.*, Vol. 92, pp. 3078–3096.
- M.G. Rahim, C.C. Goodyear, W.B. Kleijn, J. Schroeter and M.M. Sondhi (1993), "On the use of neural networks in articulatory speech synthesis", *J. Acoust. Soc. Amer.*, Vol. 93, pp. 1109–1121.
- P. Rubin, T. Baer and P. Mermelstein (1981), "An articulatory synthesizer for perceptual research", *J. Acoust. Soc. Amer.*, Vol. 70, pp. 321–328.
- E. Saltzman (1986), "Task-dynamic coordination of the speech articulators: A preliminary model", *Experimental Brain Research*, Vol. 15, pp. 129–144.
- E.L. Saltzman and J.A.S. Kelso (1987), "Skilled actions: A task-dynamic approach", *Psychological Review*, Vol. 94, pp. 84–106.
- E.L. Saltzman and K.G. Munhall (1989), "A dynamic approach to gestural patterning in speech production", *Ecological Psychology*, Vol. 14, pp. 333–382.
- M.R. Schroeder (1967), "Determination of the geometry of the human vocal tract by acoustic measurements", *J. Acoust. Soc. Amer.*, Vol. 41, pp. 1002–1010.
- J. Schroeter and M.M. Sondhi (1989), "Dynamic programming search of articulatory codebooks", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.* 89, pp. 588–591.
- J. Schroeter and M.M. Sondhi (1992), "Speech coding based on physiological models of speech production", in *Advances in Speech Signal Processing*, ed. by S. Furui and M.M. Sondhi (Marcel Dekker, New York), pp. 231–268.
- J. Schroeter, J.N. Larar and M.M. Sondhi (1987), "Speech parameter estimation using a vocal tract/cord model", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.* 87, Dallas, TX, pp. 308–311.
- J. Schroeter, P. Meyer and S. Parthasarthy (1990), "Evaluation of improved articulatory codebooks and codebook access distance measures", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.* 90, Albuquerque, NM.
- K. Shirai and T. Kobayashi (1986), "Estimating articulatory motion from speech wave", *Speech Communication*, Vol. 5, No. 2, pp. 159–170.
- K. Shirai and T. Kobayashi (1991), "Estimation of articulatory

- motion using neural networks”, *J. Phonetics*, Vol. 19, pp. 379–385.
- M.M. Sondhi and J.R. Resnick (1983), “The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis”, *J. Acoust. Soc. Amer.*, Vol. 73, pp. 985–1002.
- M.M. Sondhi and J. Schroeter (1987), “A hybrid time-frequency domain articulatory speech synthesizer”, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35, pp. 955–967.
- K.N. Stevens (1960), “Toward a model for speech recognition”, *J. Acoust. Soc. Amer.*, Vol. 32, pp. 47–55.
- K.N. Stevens (1993), “Use of acoustic data to infer articulatory movements for consonants”, *Third Seminar on Speech Production, Old Saybrook, CT*, presented.
- K.N. Stevens and A.S. House (1955), “Development of a quantitative description of vowel articulation”, *J. Acoust. Soc. Amer.*, Vol. 27, pp. 484–493.
- H. Wakita (1973), “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms”, *IEEE Trans. Audio and Electroacoust.*, Vol. 21, pp. 417–427.