Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data

George Papcun, Judith Hochberg, Timothy R. Thomas, François Laroche, Jeff Zacks, and Simon Levy

Computing and Communications Division, Mail Stop B265, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 25 June 1991; accepted for publication 10 March 1992)

This paper describes a method for inferring articulatory parameters from acoustics with a neural network trained on paired acoustic and articulatory data. An x-ray microbeam recorded the vertical movements of the lower lip, tongue tip, and tongue dorsum of three speakers saying the English stop consonants in repeated Cə syllables. A neural network was then trained to map from simultaneously recorded acoustic data to the articulatory data. To evaluate learning, acoustics from the training set were passed through the neural network. To evaluate generalization, acoustics from speakers or consonants excluded from the training set were passed through the network. The articulatory trajectories thus inferred were a good fit to the actual movements in both the learning and generalization conditions, as judged by root-mean-square error and correlation. Inferred trajectories were also matched to templates of lower lip, tongue tip, and tongue dorsum release gestures extracted from the original data. This technique correctly recognized from 94.4% to 98.9% of all gestures in the learning and cross-speaker generalization conditions, and 75% of gestures underlying consonants excluded from the training set. In addition, greater regularity was observed for movements of articulators that were critical in the formation of each consonant.

PACS numbers: 43.72.Ct, 43.72.Ne, 43.70.Bk, 43.70.Hs

INTRODUCTION

It has long been a goal of speech science to infer from acoustic data characteristics of the articulation that produced those data (the "vocal tract inverse problem"). This capability would have many clinical and technological applications in addition to its obvious utility for speech science research. A visual display of inferred articulatory movements would benefit individuals with speech disorders, who could get visual feedback on whether they correctly articulated an utterance. It would likewise benefit deaf individuals in oral therapy, and could supplement lip-reading as an aid in comprehension.[1] Inferred articulatory parameters would also make possible low bit-rate speech transmission, and could serve as the basis for fluent, connected speech recognition. From a theoretical standpoint, our approach is motivated by Browman and Goldstein's articulatory phonology (1987, 1990), according to which speech is seen as the product of temporally overlapping dynamical regimes, or gestures, each of which regulates the formation of a constriction in the vocal tract. To the extent that we can accurately infer articulation from acoustics, we can make use of inferred data in further exploring and testing the theory of gestural phonology.

Most previous research on inferring articulation from acoustics has been done within an analytic framework, in which various mathematical techniques are applied to acoustic data in order to yield area functions for the vocal tract. The earliest analytic work, and some recent work, analyzes the impulse response at the lips: the characteristics of an acoustic pulse sent into the vocal tract from an external source and reflected back (e.g., Schroeder, 1967; Mermelstein, 1967; Sondhi, 1979; Sondhi and Resnick, 1983; Milenkovic, 1984, 1987). While good results have been attained using this method, it is of limited utility, since it cannot be applied to normal speech data, whether live or recorded. In addition, the requirement that users must silently articulate into an impedance tube that is applied to their lips via a coupler probably distorts articulation, and prevents the device's being used to distinguish voiced and voiceless sounds. Other, more applicable, analytic research analyzes the acoustic signal itself, as recorded during normal speech (e.g., Atal, 1970; Wakita, 1973, 1979). This method, however, is ill-suited for nasalized or voiceless sounds. Even for those sounds that are most amenable to direct analysis, inferred area functions may be physiologically impossible or nonunique (Larar *et al.*, 1988). Kac (1966) and subsequent researchers (e.g., Protter, 1987) have also explicated the difficulties of the inverse problem in general, which suggests that the analytic approach may be unworkable *a priori*.

For both the impulse response and direct acoustic methods, an additional weakness is the difficulty of evaluating the correctness of the inferred area function. In the absence of x-ray tracings of a speech sample, there is no direct way to verify that the function inferred for the sample duplicates the function that produced it. And indirect verification—com-

paring an input sound with speech synthesized from its inferred area function—is problematic. Because the relationship of articulatory to acoustic parameters is nonunique, as mentioned above, inferred and actual functions can differ even when synthesized and actual speech do not.

Recent years have seen increasing use of articulatory-to-acoustic models of speech synthesis to finesse some of the difficulties of the analytic method. Most straightforwardly, speech can be repeatedly synthesized from hypothesized articulatory parameters that are adjusted until the synthetic speech differs minimally from the actual speech acoustics (e.g., Flanagan *et al.*, 1980; Levinson and Schmidt, 1983; Schroeter *et al.*, 1986). In the "codebook" and "sorting" methods, articulatory parameters and their corresponding synthesized acoustic parameters form a database (Larar *et al.*, 1988; Atal *et al.*, 1978). New acoustic input is analyzed by finding a similar acoustic entry in the database and outputting the corresponding articulatory parameters. In the codebook method, these parameters are then used as the initial hypothesis for analysis-by-synthesis optimization. In Atal and Rioul's neural network approach (1989), a similar database is used as a training set for a neural network. New acoustic data are put through the network in order to infer the underlying articulations. More recently, Rahim *et al.* (1991a,b) have extended this approach to include training on actual speech acoustics.

Synthesis-based methods overcome many of the difficulties of the analytic methods. Because they accept normal acoustic input, no special apparatus is required to make articulatory inferences. Because the source database can contain articulatory/acoustic pairs that span the articulatory space, there are no restrictions on what kind of speech can be addressed. Atal and Rioul's approach is especially appealing because neural networks seem ideally suited to speech analysis in their ability to make use of information in a context-sensitive fashion, interpreting acoustic energy at a particular time and frequency as evidence for different articulations depending on its temporal and spectral context. However, Atal and Rioul's method, and synthesis-based methods in general, share with the analytic methods the problem of evaluation. In addition, any weaknesses (known or unknown) of the synthesis model are passed on to the inverse technique based on it.

In this paper, we describe a neural network approach to inferring articulation that differs from prior research in a key respect: *both the acoustic data and the articulatory data in the training set are real, not synthesized.* This innovation has been made possible by the development of an x-ray microbeam that tracks movements of gold pellets attached to different articulators (Kiritani *et al.*, 1975; Nadler *et al.*, 1987; Abbs *et al.*, 1988). Since the microbeam emits a low level of radiation, one can safely x-ray speakers during extended speech samples. The microbeam data, paired with acoustic data recorded simultaneously, form the training set for the neural network. The network can be tested with additional paired data, thus solving the evaluation problem.

We inferred articulation from a basic set of speech sounds: the English stop consonants /p/, /b/, /t/, /d/, /k/, and /g/. The articulators whose movements we inferred were also a basic set: the lower lip, tongue tip, and tongue dorsum. While a neural network could be used to infer movements of more of the vocal tract—anywhere that a pellet can be placed—these three articulators provided a rudimentary representation of the oral tract and are the oral articulators most critically involved in the production of the stop consonants.

In addition to inferring articulation, we explored the possibility of recognizing speech gestures from inferred trajectories of the three articulators. While the primary purpose served by gesture recognition in this paper was that of verifying the accuracy of inferred trajectories, it was intrinsically interesting on other grounds as well. Our ability to find gestures served as a test of the construct validity of a gesture-based phonological theory such as Browman and Goldstein's articulatory phonology (1987, 1990). From a clinical perspective, gesture recognition could serve as a useful yes-or-no therapeutic measure of correct speech production, by itself or coupled with a visual display of inferred trajectories. Gestures recognized from inferred trajectories would also be an important part of an articulation-based speech recognition system.

The main finding this paper describes is that the neural network successfully inferred articulatory movements. Trajectories of inferred and actual movements over time were close to each other in Euclidean distance and had a similar shape. On the basis of these trajectories we were able to recognize the gestures underlying most consonants. This finding was robust: while inferred trajectories and the gestures recognized from them were most accurate for speech included in the training set, good results were also obtained for speech, speakers, and speech sounds not included in the training set.

Our research also sheds light on the dynamic roles that different articulators play in making the different stop consonants, a topic that the x-ray microbeam data opens up to direct study for the first time. We found that the movements of the articulator critical to the articulation of each consonant—the lower lip for /p/ and /b/, the tongue tip for /t/ and /d/, and the tongue dorsum for /k/ and /g/—were more regular than movements of the noncritical articulators. We discuss this phenomenon's effect on the neural network's performance, and its implications for our understanding of articulatory control.

I. METHOD

The overall structure of our study follows the standard for neural network studies (e.g., McClelland and Rumelhart, 1986). Data are collected and processed. From these data a training set is selected, consisting of input/output vectors. These vectors are presented to a neural network—a set of units with weighted interconnections—which "learns" a relationship between the input and output values of each vector on the basis of a series of training runs (i.e., exposure to the training set). Weights are updated after each training run according to the error function between the vector's actual output values and those predicted by the network's cur-

rent weights. When some error minimization criterion is reached, the testing phase begins. To test how well the network has learned the data presented to it, input values from the training set are put through the network, and the output values it infers are compared to the actual output values. To test the network's ability to generalize to fresh data, a new set of input/output vectors is selected from the data and tested the same way.

In our application a separate network was trained for each articulator; we will refer to a set of three articulator-specific networks as a composite network. Input/output vectors consisted of paired acoustic and articulatory data. In training, acoustic and articulatory data were presented to the composite network simultaneously. In testing, acoustic input was used to predict articulatory output, which was compared to the actual articulatory data. Data from the training set as well as new data were tested in order to evaluate both learning and generalization.

In the following sections we describe this method in more detail: the gathering and processing of data, the makeup of input/output vectors, the neural network parameters, and the measures used to evaluate inferred articulatory data. We will limit our presentation of signal processing techniques and neural network parameters to those arrived at after a lengthy optimization process. In general, these proved to be optimal in two senses: in minimizing the time needed to train the network, and in maximizing the success of generalization (Thomas *et al.*, 1992).

A. Data collection

Articulatory and acoustic data were recorded at the x-ray microbeam facility at the Waisman Center at the University of Wisconsin at Madison. Subjects were three male students, aged 20, 21, and 25. All were native speakers of English. Two were from the midwest, and one was from California. We will refer to the three subjects as 4s, 5s, and 6s, since these were the fourth, fifth, and sixth subjects we recorded at the microbeam facility.

The microbeam records articulatory data by emitting a fine x-ray beam (1-mm diameter) toward gold pellets (3-mm diameter) affixed to a subject's articulators. The pellets are sampled one at a time, with the microbeam cycling among the different pellets at a rate necessitated by the velocity of their host articulators: the faster the articulator, the higher the sampling rate. Each time a pellet is sampled, a detector on the other side of the subject determines the pellet's position from the resultant x-ray image. This is recorded as measurements (in micrometers) of vertical and horizontal coordinates with respect to the occlusal plane, with the origin being the tip of the upper central incisors. The pellet's position, velocity, and acceleration are used to aim the microbeam the next time the pellet is sampled. The acoustic signal is simultaneously recorded in digital form at a sampling rate of 10,000 Hz. (For more details, see Kiritani *et al.*, 1975; Nadler *et al.*, 1987; Abbs *et al.*, 1988.)

The data considered in this paper come from pellets affixed with dental adhesive to the middle of the lower lip along the vermilion line, and to the midline of each subject's tongue at distances of 10 and 60 mm from the tongue tip (55 mm for subject 6s' second tongue pellet). We shall refer to these as the lower lip, tongue tip, and tongue dorsum pellets, respectively. We studied the pellets' vertical movements only, since these are most relevant to the consonant closures. Sampling rates for the pellets were 90 Hz for the lower lip, 180 Hz for the tongue tip, and 90 Hz for the tongue dorsum. Microbeam measurements of a pellet placed on the nose were used to subtract head movement from the data. Other articulatory data, which we hope to use in the future, were gathered from pellets placed on the upper lip, central incisors, lower jaw, and tongue body, from a pellet sutured to the velum (subjects 5s and 6s only), from a tracing of each subject's palate, and from accelerometers taped to each subject's nose and throat (outside the glottis). For all the data used for the investigations reported in this paper, subjects were instructed to speak normally, with no other instructions as to speech rate or degree of formality.

The data consisted of 6 records for each subject, each containing articulatory and acoustic measurements for an utterance of the form CəCəCəCəCə, with all C's identical. A typical record (subject 4s's/t/ record) is illustrated in Fig. 1. Each subject produced one record with /p/, one with /b/, etc. These data were part of a larger consonant data set we gathered in Wisconsin. Each record in this larger set contained three CəCəCəCəCə utterances, using three different consonants, with a pause between each CəCəCəCəCə. We drew our data from records containing /t,l,p/, /d,v,y/, /g,b,ð/, and /k,ʒ, and s/, cutting from the longer records so as to leave a 239.7-ms silence buffer on either side (30 time-steps, as defined in Sec. I B).

Because of an error in the computer program that recorded the data, the acoustic and articulatory data were mis-

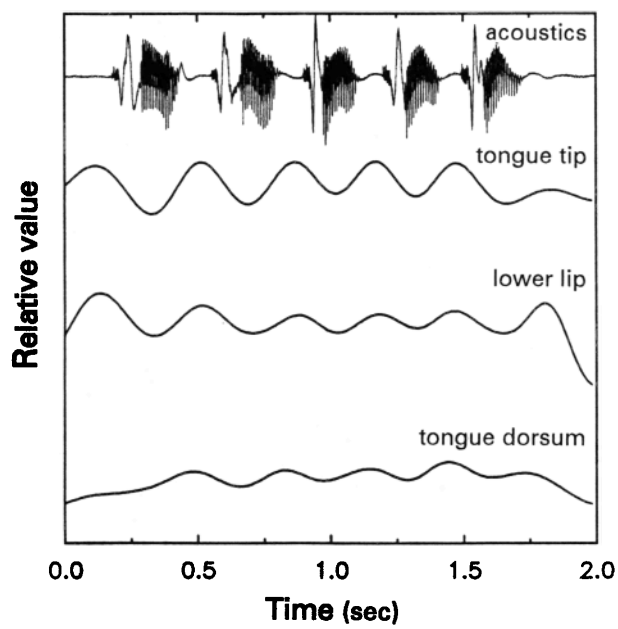

FIG. 1. Paired acoustic and articulatory data for subject 4s saying /tətətətə/.

aligned in subject 5s's /k/ record, although the articulators were correctly aligned with each other. We realigned this record "by hand," offsetting the articulatory data by small intervals until the alignment appeared similar to that of the other subjects. The eventual offset needed was 127.84 ms (16 timesteps).

B. Signal processing

The acoustic data in each record were low-pass filtered at 4000 Hz and resampled at Codec rate (8012.821 Hz) to make our research more compatible with real-world demands such as telephone communication and NeXT computer soundkit applications. The data were then adjusted to a mean of 0 to remove any dc bias and segmented into 50% overlapping 15.98-ms blocks (128 data points). This resulted in 7.99-ms timesteps between the midpoints of each pair of overlapped blocks. Each block was Welch windowed and transformed to the frequency domain with an FFT. The power spectrum produced by the FFT was converted to a decibel scale and redistributed into 18 standard bark-scale bins to represent the frequency resolution of the human auditory system. In the interest of keeping the input to the network compact, we deleted the first two bark bins (up to 200 Hz). These were observed to correlate highly with the third bin; in addition, pilot studies showed no degradation when they were removed. Finally, each record was normalized in order to eliminate chance events, where sporadic very large or small values were recorded, and to equalize loudness for different recordings of different subjects. This was accomplished by assigning the extremes 0 and 1 to the smallest and largest 0.1% of the data, and proportionate values to all other data.[2]

The signal processing also identified each timestep as containing speech or silence by comparing the average value of the 16 top bark bins to a threshold value (determined by inspection) representing background noise at the Wisconsin site. Thresholding was performed before normalization, since the average bark value of silence after normalization depended on the loudness of the speech in the record. The result of this thresholding was marked in an additional "bin" for each timestep, with silence marked with a 0, and speech with a 1.

The articulatory data in each record were synchronized to the acoustic timesteps by interpolation on a cubic spline (Press *et al.*, 1988, p. 97) and then smoothed over a 20-point window (Press *et al.*, 1988, p. 515). Data for each articulator were normalized across each speaker's six records, using the extreme values 0.1 and 0.9. These were used because 0 and 1 (the extremes of the acoustic normalization) are unsuitable for output data, being the unrealizable limits of the nonlinear sigmoid function applied to the final network output. As in the acoustic data, normalization was performed to eliminate chance events. Its most important function, however, was to equalize the range of tongue movements between speakers. Since the three subjects had different palate shapes, the same (unnormalized) vertical position of an articulator could represent a closed vocal tract for one subject but an open tract for another.

The processed data for each record, then, consisted of an acoustic vector of size $T*17$, where T is the number of timesteps in the record, each represented by 16 bark bins plus a 17th "bin" designating speech or silence, and three vectors of size T for the lower lip and the tongue tip and dorsum.

C. Input and output

Each input/output vector presented to the composite neural network had as input acoustic data from 25 timesteps (199.75 ms), which we will call a context frame, and as output the position of the three articulators at one of these timesteps, which we will call the focus timestep. The focus timestep was located 0.4 of the way into the context frame. All context frames were required to contain at least three non-silent timesteps. In training, this prevented the network from trying to map from silence to diverse output, as the tongue can be in any position during silent intervals. Such one-to-many mappings are notoriously damaging to learning in the neural network paradigm. In testing, this restriction prevented spurious output being generated from silent input. With the CəCəCəCəCə data, the only context frames rejected as containing too much silence were at the beginning and end of each record.

When a record was used as part of a training set, input/output vectors were chosen from it at 4-timestep intervals; thus only 1/4 of the potential vectors were used in training. When a record was used for testing, continuous output was produced using input from all possible vectors. Output in the testing phase was smoothed over a 10-point window.

D. Network parameters

The three articulator-specific neural networks that made up the composite network all had the same structure. There were 400 input units (25 timesteps*16 bark bins), two hidden layers of eight units each, and one output unit. Weights were initially set to random values between ± 0.8, and were updated after each training run according to a backpropagation algorithm (gradient descent plus momentum). The weight update equation was

$$\Delta w(t) = \alpha \frac{\partial E}{\partial w}(t) + \mu \Delta w(t-1),$$

where α is the learning rate, μ is a momentum factor, set to 0.3, and $(\partial E/\partial w)(t)$ is the error derivative, computed by the backpropagation algorithm. Training continued until the root mean square error between the actual and predicted output values was less than 0.08.

Code was written in FORTRAN and C specifically for this project. Our original implementation on a Sun took 3 days for each training run. After implementing the network on a Cray YMP, and optimizing it using the Cray's code optimization utilities, a typical training run now takes 40 to 50 CPU seconds.

E. Training and testing

We trained and tested the composite neural network several times, each time using a different subset of the 18

records as a training set. There were four distinct types of training regimes, as described below. The first was used to evaluate the success of learning. The other three were used to evaluate the success of generalization under a number of different conditions.

(i) Full training. In this regime, the network was trained on all 18 records, then tested on all records.

(ii) All-but-one-record training. In this regime, the network was trained on 17 records and tested on the 18th. Training and testing were repeated 18 times in a "jackknife" fashion so that each set of 17 records in turn served as training for the 18th.

(iii) Two-subject training. In this regime, the network was trained on data from two subjects (12 records) and tested on the remaining speaker (6 records). Like all-but-one-record training, this was done in a jackknife fashion so that each speaker in turn was the object of generalization.

(iv) Five-consonant training. In this regime, the network was trained on five consonants (15 records) and tested on the sixth consonant (3 records), again in a jackknife fashion.

It is interesting to compare the different generalization conditions to real-world speech recognition conditions. Speaker-dependent speech recognition involves familiarizing a system with a particular speaker, then recognizing that speaker's speech. The all-but-one-record regime was similar to this in that the network was trained on the test speaker's speech, but was different in that it was not trained on a full sample from that speaker. The two-subject regime corresponded to speaker-independent speech recognition, which involves recognizing the speech of unfamiliar speakers. Five-consonant training corresponded to a task—recognizing speech sounds completely excluded from the training set (the missing consonant)—that is seldom put to speech recognition devices, but which could have useful clinical applications in diagnosing pathological articulations.

F. Evaluation

The output from each test run of the composite neural network consisted of trajectories (over time) of the three articulators. Each record had four such sets of inferred trajectories, one from each training regime. We evaluated the accuracy of these trajectories by comparing them to the actual trajectories, using the two measures of root-mean-square (rms) error and Pearson product-moment correlation. Rms error quantified the overall distance between an inferred trajectory and the corresponding actual trajectory. It was calculated as the mean of the squared distance between the two trajectories at each timestep. Correlation quantified the similarity in shape between the trajectories regardless of magnitude—whether the actual and inferred trajectories rose and fell in synchrony.

A low rms error and a high correlation indicated an intuitive "good fit" between actual and inferred trajectories, as exemplified in the upper left quadrant of Fig. 2. A high rms error and high correlation indicated trajectories with similar shapes but discrepancies in magnitude, as in the low-

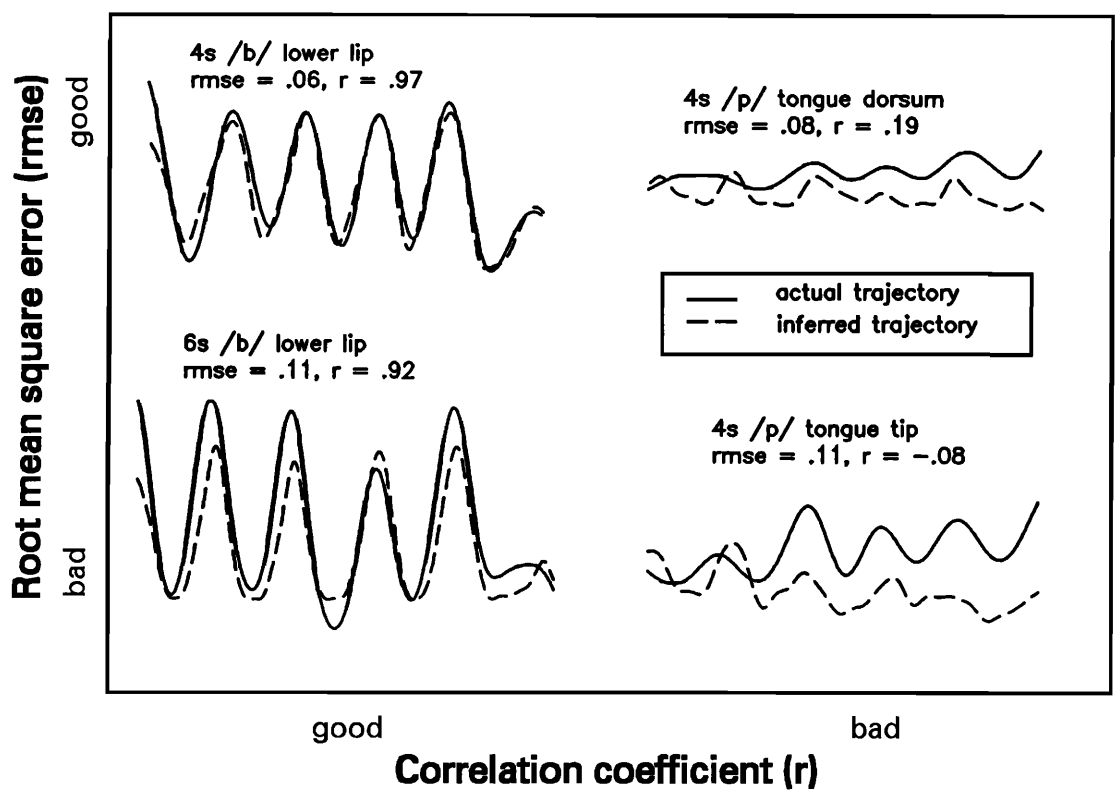

FIG. 2. Examples of actual and inferred trajectory pairs with different combinations of correlation and rms error scores.

er left quadrant of Fig. 2, where the rms error is higher than in the upper left quadrant. A low rms error and low correlation indicated trajectories that were seldom far apart, but had different shapes, as in the upper right quadrant of Fig. 2—note how the inferred trajectory often heads up when the actual trajectory heads down. A bad score on both measures—a high rms error and a low correlation—indicated problems with magnitude and shape, as in the lower right quadrant of Fig. 2. Note that the inferred trajectory contains a spurious peak at the very beginning of the record. The correct peaks are all either too high or too low, compared to the actual peaks; in addition, the first and third correct peaks are late and the fourth correct peak is flat.

While rms error and correlation were revealing, they were lacking in two respects. First, they measured the accuracy of the individual articulators as inferred from the network rather than that of the oral tract as a whole (as represented by the three pellets). Second, they did not provide an absolute assessment of the goodness of fit between actual and inferred trajectories. Correlation and rms error scores told us how accurate a trajectory was relative to other inferred trajectories, but not whether it was accurate enough to be considered a good inference on grounds beyond intuition. With over 200 timesteps in the average trajectory, the normal confidence levels for correlations became meaningless—even a small correlation was likely to be statistically significant, but that did not guarantee a good match. For example, the correlation of 0.19 shown in the upper right quadrant of Fig. 2 is statistically significant at the 0.01 level. Nonetheless, we would not want to call the (lack of) correspondence of the curves shown there a good match. Nor was there an available statistic for the rms error needed to prove accuracy.

To resolve these problems (and for reasons of intrinsic interest, as noted in the Introduction) we added a third measure of accuracy: gesture recognition. In contrast to rms error and correlation, gesture recognition provided an absolute standard of accuracy. Sufficiently accurate inferred gestures were correctly recognized; insufficiently accurate inferred gestures were not. Next, we present the details of the gesture recognition process.

From the original articulatory data, we developed templates of the movements of the three articulators during the production of bilabial, alveolar, and velar consonants. These templates covered the "release gesture" of each Cə syllable, which we defined as the interval from the consonant closure to the vowel peak or center. Templates were based on the relevant Wisconsin records for each gesture type: /p/ and /b/ records for the bilabial template, /t/ and /d/ for the alveolar, /k/ and /g/ for the velar. As each record contained five repetitions of each syllable, the three speakers provided a total of 30 syllables for the construction of each template (i.e., 3 speakers*5 repetitions*2 voicing conditions).

The first step in making a template was to identify the boundaries of the release gestures of the 30 contributing syllables, which we defined as the interval from the consonant closure to the vowel peak, or center. Gestural onsets were found by examining the trajectory of the "critical articulator" involved in a consonant's production: the lower lip for the bilabials, the tongue tip for the alveolars, and the tongue dorsum for the velars. In each case, we smoothed the critical articulator trajectory over a 10-timestep window, then used a peak-picking algorithm to find the five maxima preceding the five vowels in each record. Offsets were found by taking the point 15 timesteps after the gesture's onset, this being the minimum release gesture length in our data.[3] Once each gesture's boundaries had been determined, the 30 exemplars of each gesture were averaged to form a template. The resulting templates are shown in Fig. 3.

To perform the actual gesture recognition from the inferred trajectories, a similar technique was used in locating the boundaries of the inferred release gestures. Each gesture in the inferred trajectories was compared to the bilabial, alveolar, and velar templates. Gestural boundaries were defined according to the template currently being matched. For example, in comparing a syllable to the bilabial template, the gestural onset was defined by the lower lip maximum (after smoothing) preceding the vowel peak; the offset was again 15 timesteps after the onset. The distance between each articulator's inferred and template values was then measured at each timestep, and the root-mean-square error of these distances was calculated as an error measure for the degree of match between the inferred trajectory and that template.

Once this comparison had been carried out for all three templates, the template that yielded the lowest error score was selected as the best match for that gesture. Thus we were able to say whether a /bə/ syllable was correctly recognized as containing a bilabial gesture, a /tə/ syllable an alveolar, etc. As a test of the gesture recognition technique, we attempted to identify the 90 gestures in the actual trajectories. All but one were correctly recognized.

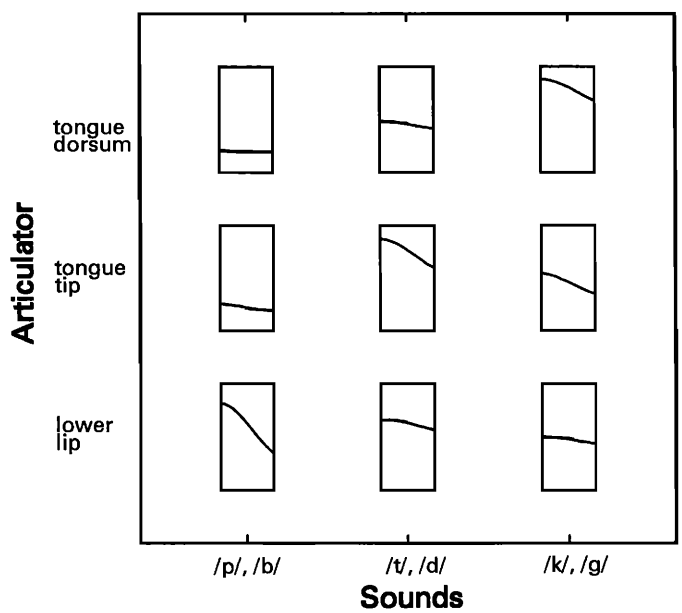

FIG. 3. Articulatory templates used in the gesture recognition measure. Each template describes the movement of a single articulator during a bilabial (/p/, /b/), alveolar (/t/, /d/), or velar (/k/, /g/) release gesture, and extends 15 timesteps (119.85 ms) from the consonant closure.

TABLE I. Average correlation coefficients (with s.d.) between actual trajectories and trajectories inferred in the full training regime.

Sound	Lower lip	Tongue tip	Tongue dorsum
p	0.90 (0.02)	0.30 (0.33)	0.19 (0.12)
b	0.94 (0.03)	0.70 (0.10)	0.44 (0.04)
t	0.78 (0.15)	0.88 (0.04)	0.65 (0.18)
d	0.75 (0.19)	0.89 (0.08)	0.26 (0.25)
k	0.59 (0.22)	0.73 (0.20)	0.90 (0.07)
g	0.53 (0.22)	0.72 (0.11)	0.90 (0.01)

TABLE II. Average root mean square error (with s.d.) between actual trajectories and trajectories inferred in the full training regime. Trajectories are normalized between 0.1 and 0.9 for each speaker and articulator between records.

Sound	Lower lip	Tongue tip	Tongue dorsum
p	0.11 (0.01)	0.11 (0.04)	0.07 (0.02)
b	0.10 (0.03)	0.08 (0.03)	0.06 (0.00)
t	0.08 (0.01)	0.08 (0.01)	0.06 (0.00)
d	0.06 (0.01)	0.07 (0.01)	0.09 (0.02)
k	0.09 (0.02)	0.08 (0.02)	0.10 (0.04)
g	0.09 (0.02)	0.10 (0.02)	0.10 (0.03)

II. RESULTS

A. Learning

The results of the full training regime showed that the composite neural network successfully learned the relationship between acoustics and articulation. Let us first consider the accuracy of the individual trajectories, as measured by correlation coefficient and rms error. As shown in Table I, average correlations ranged from 0.19 to 0.94. The highest correlations, as well as the lowest standard deviations, were consistently found for trajectories of the "critical articulator" for each consonant type: the lower lip for labials /p/ and /b/, the tongue tip for alveolars /t/ and /d/, and the tongue dorsum for velars /k/ and /g/. This result is shown graphically in Fig. 4, which combines results for the voiced and voiceless consonants.

The opposite pattern was found for rms error. As shown in Table II, error was low in general, with averages ranging from 0.06 to 0.11; this was to be expected, since the training termination criterion was an rms error of 0.08 for the training set as a whole. But rms error was generally higher for the critical articulators than for the noncritical articulators, as shown graphically in Fig. 5. The sole exception was /d/, for which the rms error was higher for the tongue dorsum than the tongue tip.

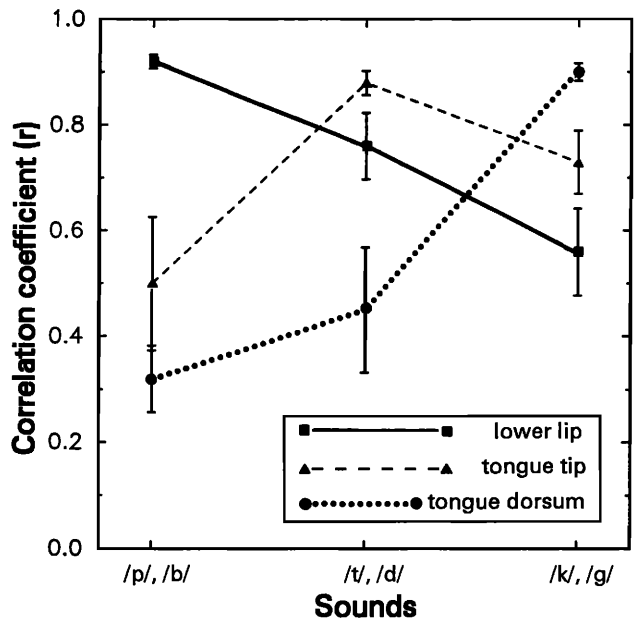

FIG. 4. Average correlation coefficient (r) between actual trajectories and trajectories inferred in the full training regime. Error bars represent the standard error of the mean.

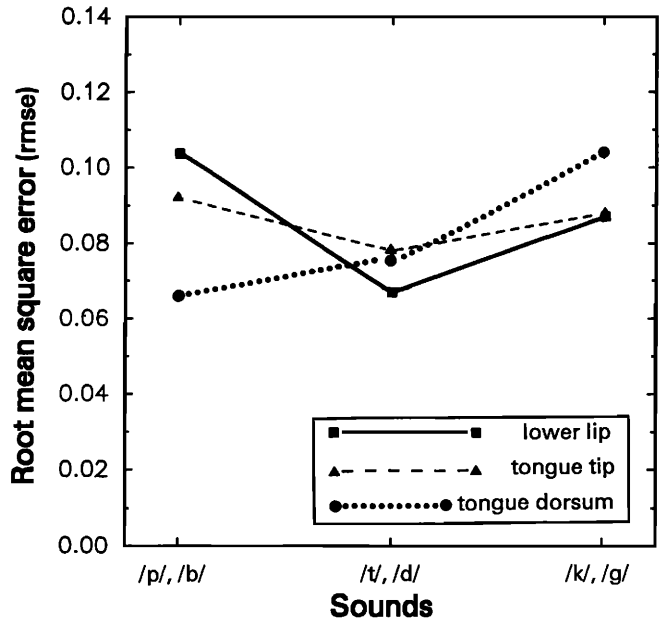

FIG. 5. Average root-mean-squared error (rmse) between actual trajectories and trajectories inferred in the full training regime. Error bars are omitted because of excessive overlap.

In Sec. II B we wil investigate the cause of this "critical articulator" phenomenon: that critical articulators had better correlations but worse rms errors than noncritical articulators. Our present concern is what this phenomenon tells us about the accuracy of the inferred trajectories. For each consonant type in our data, the key movement involved was the raising and lowering of the critical articulator; i.e., the shape of the critical articulator trajectory. For the noncritical articulators, in contrast, position was key: these articulators needed only to remain at the appropriate position for the schwa and to avoid forming a closure. Our data accordingly exhibited a smaller range of movement for noncritical articulators than for critical articulators (Table III). The high correlations for the critical articulators showed that their shapes were learned well, while the low rms errors for the noncritical articulators showed that their positions were inferred accurately. These two patterns are illustrated in the lower left and upper right quadrants of Fig. 2. Thus the neural network captured the ups and downs of the critical articulator and the position of the noncritical articulators, which were the essence of the consonant articulations.

This conclusion is buttressed by the results of applying our gesture recognition technique to the inferred trajectories: as in the actual data, all but one of the 90 gestures were correctly recognized.

B. Generalization

The neural network, when trained on only part of the data, successfully generalized to the rest of the data, as judged by the results of the other training regimes. As in the previous section, let us first consider the accuracy of the individual trajectories, as measured by correlation coefficient and rms error (Figs. 6 and 7). For these measures, the accuracy of inferred trajectories was linked to the number of records in the training set. As this dropped from 18 (full training) to 12 (two-subject training), correlations declined and rms error increased for both critical and noncritical articulators. Using the analysis of variance F statistic, $F = 11.43$, $p < 0.007$ for correlation, $F = 12.45$, and $p < 0.005$ for rms error, thus showing that the reduction in the number of training records produces significantly poorer accuracy of the inferred trajectories. Even in the worst training regime, average correlations never dropped below 0.73 for critical articulators and 0.39 for noncritical articulators; rms error never rose over 0.16 for critical articulators and 0.14 for noncritical articulators. The critical articulator phenomenon noted in the full training regime held in the generalization regimes as well: critical articulators had significantly better correlations ($F = 48.5$, $p < 0.02$) but significantly worse rms errors ($F = 48.75$, $p < 0.02$) than non-critical articulators.

TABLE III. Average range of normalized movement (with s.d.) for critical and noncritical articulators. Range is calculated as the difference between the maximum and minimum displacement within each trajectory.

Sounds	Critical	Noncritical
p,b	0.77 (0.06)	0.27 (0.09)
t,d	0.55 (0.13)	0.48 (0.13)
k,g	0.58 (0.10)	0.44 (0.10)

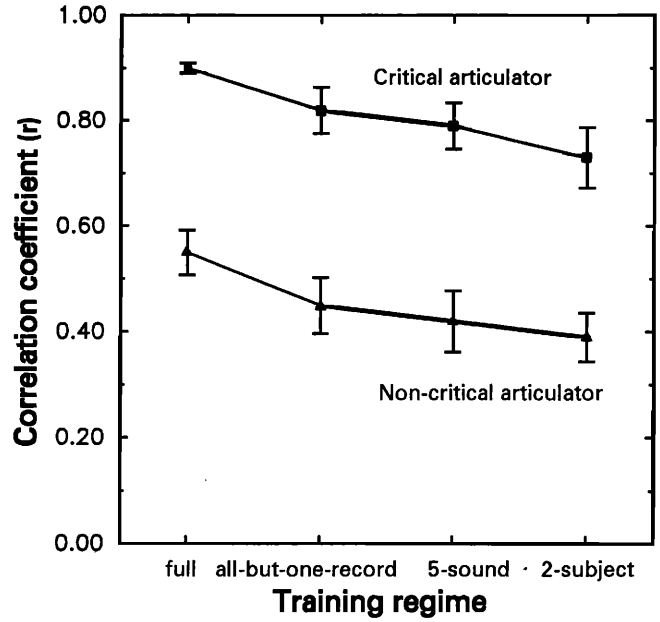

FIG. 6. Average correlation coefficient (r) between actual and inferred trajectories of critical and noncritical articulators in the four training regimes. Error bars represent the standard error of the mean.

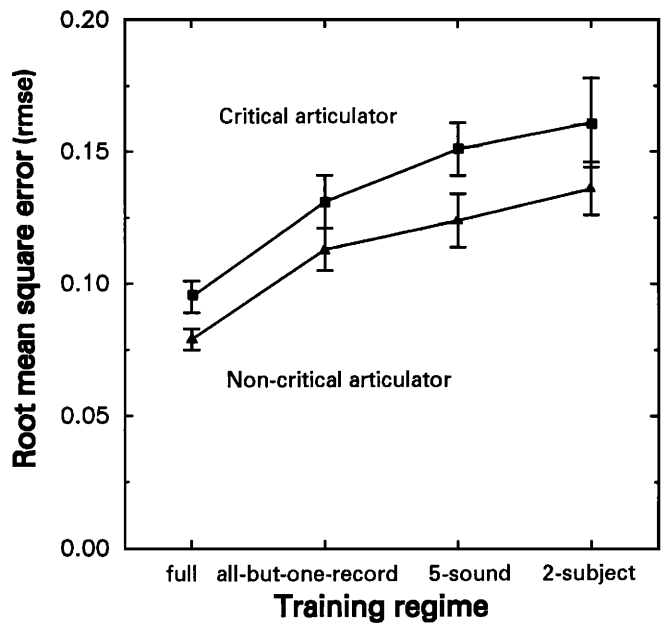

FIG. 7. Average root-mean-squared error (rmse) between actual and inferred trajectories of critical and noncritical articulators in the four training regimes. Error bars represent the standard error of the mean.

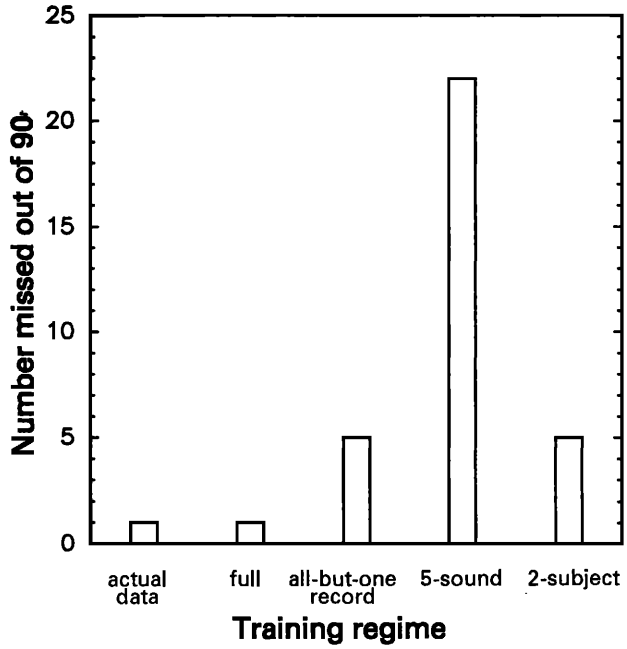

FIG. 8. Number of gestures (out of 90) misidentified in the actual data and in the four training regimes.

The gesture recognition results further demonstrated the network's ability to generalize (Fig. 8). In the all-but-one-record and two-subject regimes, only five gestures (out of 90) failed to be correctly recognized. In the five-consonant regime 22 gestures out of 90 were misrecognized. We consider this a good result since it means that we were able to correctly recognize three out of four gestures even when the test consonant was excluded from the training set; recall that this is a task seldom put to existing speech recognition devices (Sec. I E). However, we were struck by the disparity between the five-consonant regime's relatively poor showing in the gesture recognition measure, compared to its relatively good showing in the correlation and rms error measures.

We hypothesized that the answer to this anomaly lay in the type of information lost when a single consonant was removed from the training set. Since our data consisted of paired voiced and voiceless consonants, removing one consonant meant that the network was exposed to its voiced or voiceless counterpart as well as the other four consonants. While the articulatory trajectories of voiced and voiceless consonant pairs (e.g., /p/ and /b/) were similar, the voicing difference was manifest in the acoustics preceding the vowel onset—that is, in the portion of the acoustic signal corresponding to the release gesture of the consonant. Training on only one member of a consonant pair, then, should affect the network's ability to infer the release gesture of the other member. A network exposed to /p/ but not /b/ (as well as the alveolar and velar consonants), for example, would be unfamiliar with the (voiced) acoustics of the /b/'s release gesture, and should have difficulty inferring that part of the trajectory. Since the gesture recognition process dealt exclusively with the release gesture, it was sensitive to this weakness, in contrast to correlation and rms error, which were based on trajectories of entire records.

Indeed, in the five-consonant regime rms errors during release gestures were on the average greater (worse) than during the entire records (0.134 vs 0.128). This difference explains the anomalous behavior of the five-consonant and two-subject regimes in Fig. 8. While trajectories inferred in the five-consonant regime were more accurate overall than those inferred in the two-subject regime (as shown in Fig. 7), their release gestures were less accurate (0.134 for the 5 consonant regime release versus 0.128 for the two-subject regime release). We conclude that inaccuracy of the five-consonant regime release gestures was responsible for the larger numbers of gestures incorrectly recognized in that regime.

A final aspect of our results to note is the different types of errors made in gesture recognition. As summarized in Table IV, most errors confused velars and alveolars. Interestingly, this pattern of confusion is found in human perception of voiced but not voiceless consonants, voiceless confusions being mainly between /k/ and /p/ (Miller and Nicely, 1955, p. 347). In our data, velar/alveolar confusions predominated for both voiced and voiceless consonants.

C. The critical articulator phenomenon

In this section we investigate the cause of the "critical articulator" phenomenon first noted in the neural network's learning behavior, and seen again in its generalization behavior. This phenomenon was really two phenomena: on average, critical articulators had a better correlation coefficient, but a worse rms error, than noncritical articulators. (We remind the reader that by "critical articulator" we mean the articulator most crucially involved in a consonant's production: the lower lip for bilabials, the tongue tip for alveolars, and the tongue dorsum for velars.)

TABLE IV. Error type frequencies in the gesture recognition task for the actual data and the four training regimes.

Regime	Velar→alveolar	Alveolar→velar	Velar→bilabial	Bilabial→velar	Total
actual	0	0	0	1	1
full training	1	0	0	0	1
all-but-one-record	3	2	0	0	5
two-subject	3	1	1	0	5
five-consonant	14	8	0	0	22
total	21	11	1	1	34

We hypothesized that the smaller of these effects, the poorer rms errors of the critical articulators, was a natural consequence of the greater range of movement of these articulators (recall Table III). To test this hypothesis, we renormalized the articulatory data within each record rather than across records, then re-trained the network as in the full training regime. This renormalization equalized the range of movement for each articulator between critical and noncritical uses. When the network was thus retrained, the rms error effect vanished, indicating that the greater range of the critical articulator trajectories was indeed its source.

The correlation effect was more dramatic, with critical articulator correlations almost twice as strong as noncritical articulator correlations. As possible causes for this effect we investigated the two most salient characteristics of the critical articulator trajectories: their range and regularity. As noted earlier, critical articulators had a larger range than noncritical articulators. It was also apparent that their movements were less variable (compare the left- and right-hand trajectories in Fig. 2). To quantify this difference we compared peak-to-peak trajectories of critical and noncritical articulators for each of the sound types. For example, for each labial consonant record we first found the five timesteps corresponding to the peaks of the lower lip trajectory, then spliced out the four "peak-to-peak" intervals between these timesteps for the lower lip, tongue tip, and tongue dorsum trajectories. The intervals thus selected were normalized in the time dimension to 20 equidistant timesteps, and normalized in the vertical dimension within each record (taking into account only data between the first and last consonant peaks). This procedure yielded 24 intervals for each articulator and sound type (3 speakers*2 consonants*4 peak-to-peak intervals). For each set of interval trajectories we computed the mean trajectory and the length of the variance vector. A comparison of these interval trajectories showed critical articulators to be less variable in their movements than noncritical articulators (Fig. 9). The critical articulator is, by definition, responsible for the articulation of each consonant. Accordingly, critical articulator movements should be more constrained, and noncritical articulations less constrained by segmental considerations, and therefore freer to vary.

Indirect experimental evidence linking articulatory roles and variability comes from Recasens' work on coarticulation in Catalan. Recasens observed that the more an articulator was involved in producing a consonant, the less susceptible it was to coarticulatory influences from adjacent vowels. Thus the consonant series [n], [ʎ], [ɲ], [j] exhibited both an increase in dorsopalatal contact and a decrease in the tongue dorsum's susceptibility to coarticulation from adjacent vowels, as measured from electropalatographic data (Recasens, 1984). Likewise, consonants whose articulation imposed a greater degree of constraint on the vocal tract (e.g., labiovelars) were less susceptible to coarticulation than were less constrained consonants (e.g., labials or velars), as judged by $F2$ variability (Recasens, 1985).

Given that noncritical articulators were freer to vary, it remains to be determined why they did so. One possibility is that variability was an inter-speaker phenomenon: that each

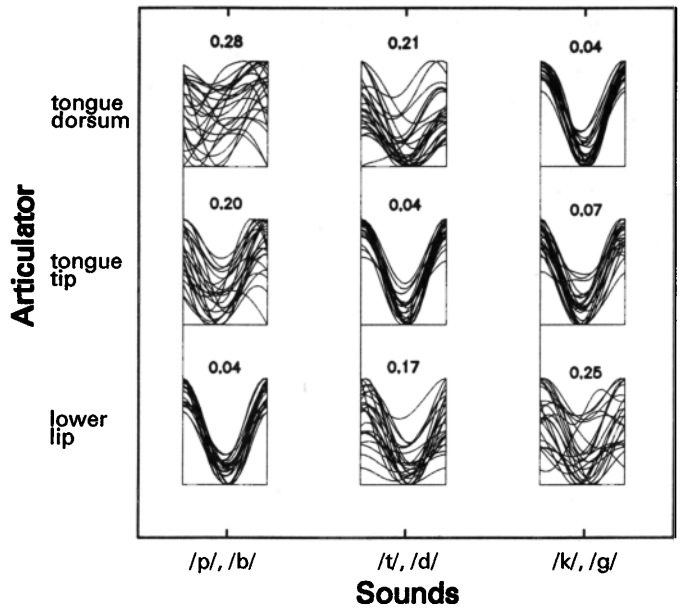

FIG. 9. Peak-to-peak trajectories of vertical movement of the lower lip, tongue tip, and tongue dorsum for the six consonants. Each section of the figure shows data from all three subjects and the length of the variance vector.

speaker had acquired idiosyncratic patterns of noncritical movement, though sharing patterns of critical articulator movement with other speakers. A second possibility is that intra-speaker variability—variation among different utterances by the same speaker—was caused by principled differences among the utterances, such as differences in phrasal stress. Finally, intra-speaker variability could be considered "noise." Relatively unconstrained by the consonantal gesture, and lacking the need to enact a marked vowel gesture (since all vowels in our study were schwa), the noncritical articulators may simply have exhibited nonsystematic movement.

From the neural network perspective, the greater range and lesser variability of the critical articulator trajectories could both lead to higher correlations. If a curve moves dramatically up and down, and if inferred data at individual timesteps are close to the actual data (as a result of rms error training), the shape of the inferred curve will be similar to that of the actual curve, resulting in a high correlation. If the curve shows a smaller range of movement, inferred data can be close to actual data at individual timesteps without a good overall match to the curve shape, resulting in a lower correlation. Thus the network should be better able to infer the vigorous movements of the critical articulators than the smaller movements of the noncritical articulators. A neural network is also less able to infer variable patterns than regular ones: if, in training, identical input maps to distinct output values at two different times, then in testing the network will have no basis for determining which treatment to give to similar input.

To test whether differences in range and inter-speaker variability were the source of the higher correlations for critical articulators, we re-trained the network so as to elimi-

nate these two factors. To eliminate range as a factor, we renormalized the articulatory data within each record, then trained and tested the network on all records, as described above. To eliminate inter-speaker variability as a factor, we trained, then tested, the neural network on each of the three subjects separately. Intra-speaker variability could not be factored out. Instead, we simultaneously factored out both range and inter-speaker variability by normalizing within each record and training within each subject, reasoning that any remaining differences between critical and noncritical articulators could be attributed to intra-speaker variability.

These tests showed that the range and intra-speaker regularity of the critical articulators, though not inter-speaker variability, contributed to the critical articulator correlation effect. As shown in Fig. 10, critical articulator correlations were higher than noncritical correlations in all three test regimes as well as the original full training regime ($F = 44.77, p < 0.022$); this effect was found to different degrees in the different regimes, as shown by a significant interaction between training regime and articulator type (critical versus noncritical: $F = 16.76, p < 0.003$). Training and testing within speakers increased rather than decreased the effect, indicating that inter-speaker variability was not a factor in the higher critical articulator correlations. Normalizing within each record reduced the effect somewhat by improving the noncritical articulator correlations, showing that range was a factor. The importance of intra-speaker variability was confirmed by the fact that a substantial effect remained when the network was trained within speakers after normalizing within records, thus removing both range and inter-subject variability as factors. Determining the cause of this intra-speaker variability is beyond the scope of this paper.

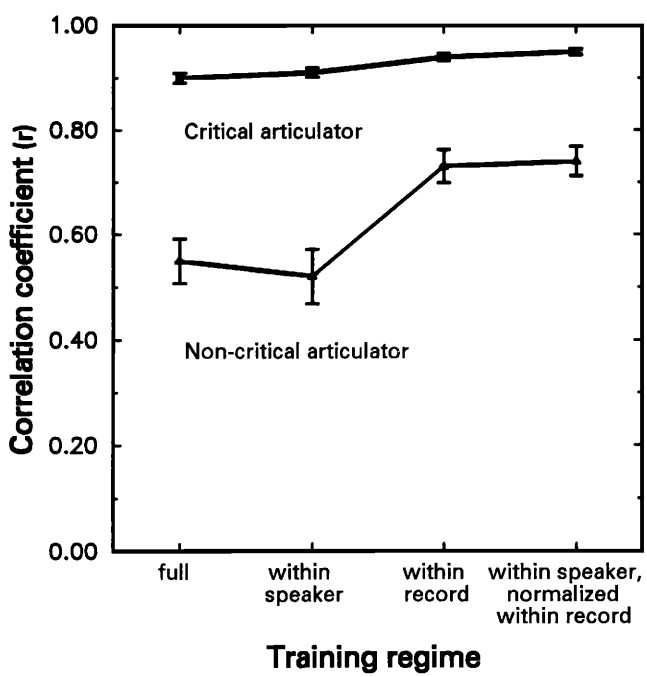

FIG. 10. Differences between average correlations for critical and noncritical articulators for the original (fully trained) network, a network trained and tested on each subject separately, a network trained after normalizing within each record, and a network trained and tested on each subject separately after normalizing within each record.

This confirmation of intra-speaker variability is relevant not only to our research, but also to our understanding of articulation and of motor control. Recasens's research has established that the position of noncritical articulators is freer to vary than that of critical articulators. We have shown that freedom to vary applies to articulatory paths as well.

III. SUMMARY AND CONCLUSIONS

We have described an approach to inferring articulatory trajectories that draws on advances in articulatory measurement technology (the x-ray microbeam) and nonlinear mapping methods (neural networks). Our results show that this approach holds great promise. After training on the English stop consonants the neural network learned the essence of the corresponding articulations, as judged by correlation and rms error between inferred and actual trajectories. That is, the inferred trajectories captured the ups and downs of the articulator most crucially involved in each consonant's articulation (the "critical articulator"), and the position of the other, "noncritical" articulators. Correlation and rms error were also strong when the network trained on only part of the data, then generalized to consonants from the training set spoken by familiar speakers, or to speakers or consonants absent in training.

We also succeeded in recognizing gestures, which we defined as dynamic patterns of articulation across the three articulators. Comparing release gestures (consonant openings) in inferred trajectories to templates derived from the actual data, we were able to recognize all but one gesture out of 90 after learning the entire data set (just as in the original data), and all but five when generalizing to sounds from the training set spoken by familiar or unfamiliar speakers. We were also able to recognize 75% of gestures underlying consonants excluded from the training set.

Finally, we observed consistent differences between critical and noncritical articulators. Compared to noncritical articulators, movements of critical articulators had a greater range and were less variable. These differences led to relatively good correlations and relatively poor rms errors for the critical articulators. The consistency of critical articulations between speakers and utterances, coupled with our ability to recognize gestures, supports Browman and Goldstein's view of speech as composed of consistent dynamical regimes.

In the future our main task will be to extend the applicability of our method. This will entail widening our training set to include more articulators, sounds, contexts, and dimensions (horizontal as well as vertical). We have already obtained further data from the Wisconsin microbeam facility for this purpose. We may also try other types of signal processing—perhaps a cochlear model, or different types of normalization. We have also developed a NeXT computer interface which lets a user record new speech, signal process it, and put it through a trained network downloaded from the Cray. The acoustics and inferred trajectories are dis-

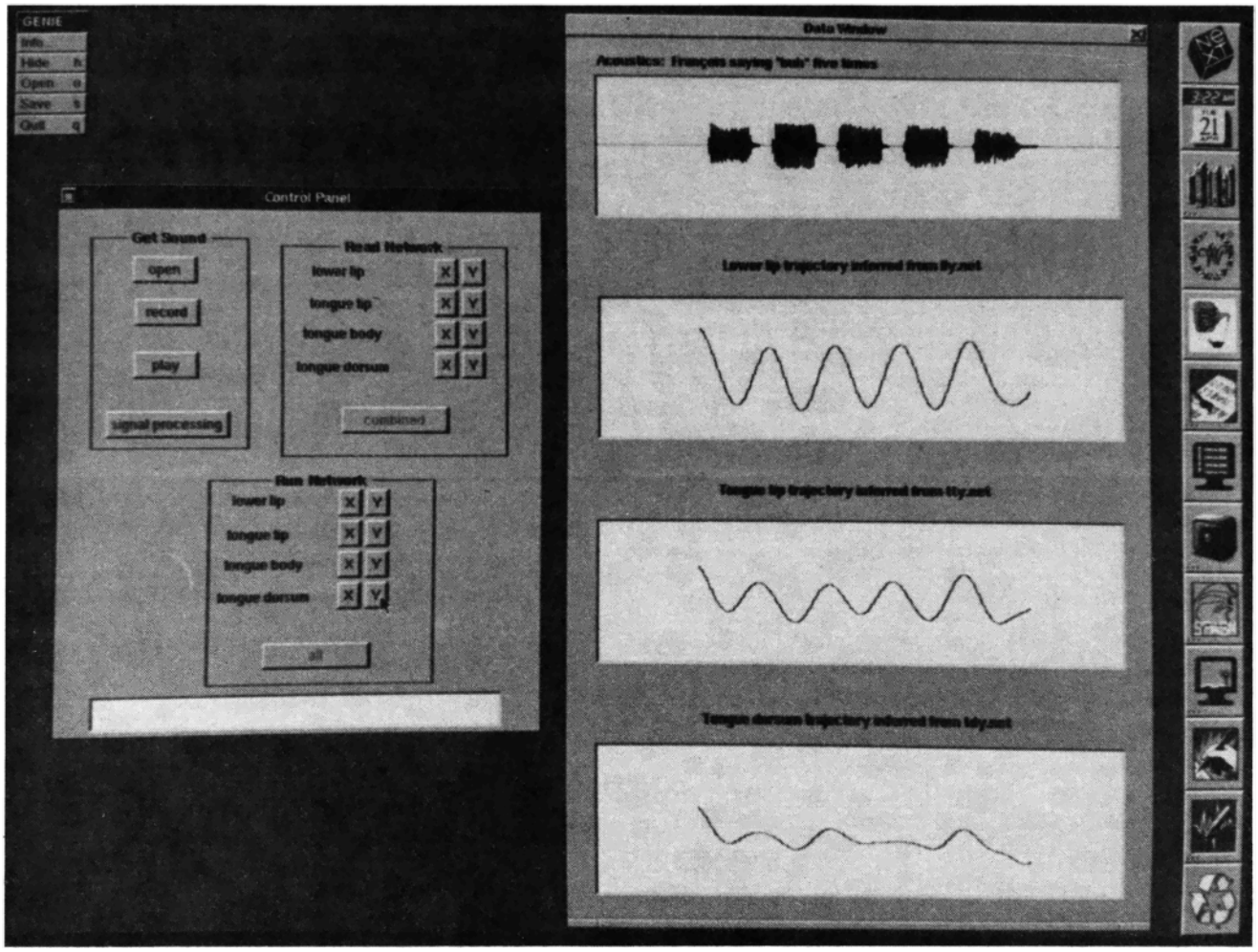

FIG. 11. Screen dump from our NeXT implementation, showing the acoustics and inferred lower lip, tongue tip, and tongue dorsum trajectories of F.L. saying /bəbəbəbəbə/.

played on the screen, as shown in Fig. 11, a screen dump from the NeXT system.

A long-term goal is to use inferred trajectories, and the gestures identified from them, as the basis for speech recognition. X-ray microbeam data suggest that gestures, unlike higher-level units such as phonemes, are relatively invariant in different contextual and speech rate conditions (Fujimura, 1986; Browman and Goldstein, 1990). For example, when the phrase *perfect memory* is spoken rapidly, the final tongue tip closing gesture of *perfect* is overlapped by the bilabial closing gesture of *memory*, but is otherwise similar to the gesture produced when *perfect* is spoken in isolation. In addition, our own data show that critical articulations are consistent among speakers who differ in voice quality, dialect, and speech rate (Fig. 9). We believe that the representational power of gestures will enable us to better deal with connected speech and different speakers, the two Achilles heels of existing speech recognition technologies.

ACKNOWLEDGMENTS

This research was performed under DOE Contract No. W-7405-ENG-36 with the University of California, and also supported in part by a grant from U. S. West Advanced Technologies. The Wisconsin Microbeam Facility is supported by NIH grant DC00162, "Speech Movement Research with an x-ray Microbeam." Carl Johnson and Bob Nadler of the Microbeam Facility were extremely helpful before, during, and after the data collection. We thank Catherine Browman and Louis Goldstein for theoretical inspiration and exceedingly valuable discussion. We would also like to thank the following students for their ideas and hard work: Tony Brewster, Susan Commisso, Barry Guinn, Darron Lockett, Clintonia Patterson, Kennan Shelton, and Mychael Vo.

[1]The idea of the lip-reading aid application was suggested to us by Hynek Hermansky, U. S. West Advanced Technologies (personal communication).

[2]Normalization took place within the full records (the original records with three CəCəCəCəCə utterances each) before our shorter records were selected from them.

[3]We tried two other ways of identifying gesture offsets. Finding offsets from critical articulator trajectory minima proved unworkable because trajectories tended to trail off downward at the end of each record rather than forming a clean "pit" for the last vowel. Identifying offsets as timesteps

with maximum acoustic energy was workable, but produced results inferior to those reported here.

Abbs, J.H., Nadler, R. D., and Fujimura, O. (**1988**). "X-Ray microbeams track the shape of speech," SOMA: Eng. Human Body **2**, 29–34.

Atal, B. S. (**1970**). "Determination of the vocal-tract shape directly from the speech wave," J. Acoust. Soc. Am. Suppl. 1 **47**, S65.

Atal, B. S. and Rioul, O. (**1989**). "Neural networks for estimating articulatory positions from speech," J. Acoust. Soc. Am. Suppl. 1 **86**, S67.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am. **63**, 1535–1555.

Browman C. and Goldstein, L. (**1987**). "Towards an articulatory phonology," Phonology Yearbook **3**, 219–252.

Browman, C., and Goldstein, L. (**1990**). "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech,* edited by J. Kingson and M. Beckman (Cambridge U.P., Cambridge), pp. 341–376.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (**1980**). "Signal models for low bit-rate coding of speech," J. Acoust. Soc. Am. **68**, 780–791.

Fujimura, O. (**1986**). "Relative invariance of articulatory movements: An iceberg model," in *Invariance and Variability in Speech*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 226–242.

Kiritani, S., Itoh, K., and Fujimura, O. (**1975**). "Tongue-pellet tracking by a computer controlled X-ray microbeam system," J. Acoust. Soc. Am. **57**, 1516–1520.

Kac, M. (**1966**). "Can one hear the shape of a drum?" Am. Math. Mon. **73**, 1–23.

Larar, J. N., Schroeter, J., and Sondhi, M. M. (**1988**). "Vector quantization of the articulatory space," IEEE Trans. Acoust. Speech Signal Process. **ASSP-36**, 1812–1818.

Levinson, S. E., and Schmidt, C. E. (**1983**). "Adaptive computation of articulatory parameters from the speech signal," J. Acoust. Soc. Am. **74**, 1145–1154.

McClelland, J. L., and Rumelhart, D. E. (**1986**). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models* (MIT Press, Cambridge).

Mermelstein, P. (**1967**). "Determination of the vocal-tract shape from measured formant frequencies," J. Acoust. Soc. Am. **41**, 1283–1294.

Milenkovic, P. (**1984**). "Vocal tract area functions from two-point acoustic measurements with formant frequency constraints," IEEE Trans. Acoust. Speech Signal Process. **ASSP-32**, 1122–1135.

Milenkovic, P. (**1987**). "Acoustic tube reconstruction from noncausal excitation," IEEE Trans. Acoust. Speech Signal Process. **ASSP-35**, 1089–1100.

Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Nadler, R. D., Abbs, J. H., and Fujimura, O. (**1987**). "Speech movement research using the new X-ray microbeam system," in *Proceedings of the XIth International Congress of Phonetic Sciences* (Academy of Sciences of Estonian Soviet Socialist Republic, Tallinn, Estonia), Vol. 1, pp. 221–224.

Press, W., Flanner, B., Teukolsky, S., and Vetterling, W. (**1988**). *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge U.P., Cambridge).

Protter, M. H. (**1987**). "Can one hear the shape of a drum? revisited," SIAM Rev. **29**, 185–197.

Rahim, M. G., Kleijn, W. B., and Schroeter, J. (**1991a**). "Neural networks in articulatory speech analysis/synthesis," J. Acoust. Soc. Am. **89**, 1892(A).

Rahim, M. G., Kleijn, W. B., Schroeter, J., and Goodyear, C. C. (**1991b**). "Acoustic to articulatory parameter mapping using an assembly of neural networks," IEEE-ICASSP, 485–488.

Recasens, D. (**1984**). "V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study," J. Phonet. **12**, 61–73.

Recasens, D. (**1985**). "Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences," Lang. Speech **28**, 97–114.

Schroeder, M. R. (**1967**). "Determination of the geometry of the human vocal tract by acoustic measurements," J. Acoust. Soc. Am. **41**, 902–1010.

Schroeter, J., Larar, J., and Sondhi, M. M. (**1986**). "Speech analysis and synthesis using a vocal tract/cord model," J. Acoust. Soc. Am. Suppl. 1 **80**, S19.

Sondhi, M. M. (**1979**). "Estimation of vocal-tract areas: The need for acoustical measurements," IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**, 268–273.

Sondhi, M. M., and Resnick, J. R. (**1983**). "The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis," J. Acoust. Soc. Am. **73**, 985–1002.

Thomas, T. R., Papcun, G., and Guinn, B. N. (**1992**). "Mapping from speech acoustics to tongue dorsum movement: An application of a multilayer perceptron," in *1991 Lectures in Complex Systems: SFI Studies in the Sciences of Complexity*, edited by D. Stein and N. Nadel (Addison-Wesley, Reading, MA), pp. 139–158.

Wakita, H. (**1973**). "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," IEEE Trans. Audio Electroacoust. **21**, 417–427.

Wakita, H. (**1979**). "Estimation of vocal-tract shapes from acoustical analysis from the speech wave: The state of the art," IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**, 281–285.