A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters

Christopher T. Kello[a)]
Department of Psychology, George Mason University, Fairfax, Virginia 22030

David C. Plaut
Department of Psychology, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

(Received 2 October 2003; revised 24 February 2004; accepted 1 March 2004)

Three neural network models were trained on the forward mapping from articulatory positions to acoustic outputs for a single speaker of the Edinburgh multi-channel articulatory speech database. The model parameters (i.e., connection weights) were learned via the backpropagation of error signals generated by the difference between acoustic outputs of the models, and their acoustic targets. Efficacy of the trained models was assessed by subjecting the models' acoustic outputs to speech intelligibility tests. The results of these tests showed that enough phonetic information was captured by the models to support rates of word identification as high as 84%, approaching an identification rate of 92% for the actual target stimuli. These forward models could serve as one component of a data-driven articulatory synthesizer. The models also provide the first step toward building a model of spoken word acquisition and phonological development trained on real speech.
 [DOI: 10.1121/1.1715112]

PACS numbers: 43.70.Bk, 43.72.Ja, 43.70.Ep, 43.70.Jt [AL] Pages: 2354–2364

I. INTRODUCTION

A necessary component of any complete model of speech acquisition or speech production is the physical relationship between the shape of the vocal tract, and the acoustic energy emitted from the vocal tract. This relationship is often referred to as the *forward* mapping from articulatory states to acoustic outputs, whereas the *inverse* mapping would recover articulatory states from the speech signal (Jordan and Rumelhart, 1992). The forward mapping is integral to speech production because the primary proximal stimulus used by the listener is the acoustic speech signal. Therefore, to produce comprehensible speech, the talker must somehow take into account the forward mapping from articulatory commands to acoustic outputs.

The articulatory-acoustic mapping has been studied primarily for two purposes. One is to better understand how speech is perceived and produced by humans (e.g., Rubin, Baer, and Mermelstein, 1981), and the other is to develop articulatory-based techniques for automatic speech recognition (e.g., Blackburn and Young, 2000a) and speech synthesis (e.g., Greenwood, Goodyear, and Martin, 1992). In the service of these purposes, computational models have been developed to simulate the forward mapping from articulation to acoustics (e.g., Baer *et al.*, 1991; Beautemps, Badin, and Laboissiere, 1995). These forward models have been based upon articulatory and acoustic dimensions that are known to convey phonetic information, and upon physical principles of the vocal tract. For instance, place of contact between the tongue and the upper surface of the oral cavity is an articulatory dimension known to play a role in distinguishing some consonants from each other (Ladefoged, 1993). Formant frequencies are acoustic dimensions known to play a role in distinguishing vowels from each other. In forward models of the articulatory-acoustic mapping, functions have been derived in order to relate these and other articulatory and acoustic dimensions to a physical model of the vocal tract (e.g., Baer *et al.*, 1991; Goodyear, 2000). Models of the vocal tract used for this purpose are commonly divided into a source of acoustic energy, and a filter through which the source is passed. One of the best known examples is the Kelly and Lochbaum (1962) model of the vocal tract in which the filter is modeled as series of tubes with varying lengths and diameters.

Most forward models developed thus far can be thought of as *theory-driven* because they are, in large part, derived from physical principles of the vocal tract (for an exception in automatic speech recognition, see Blackburn and Young, 2000a). These theory-driven models have served as valuable research tools for relating the underlying theories to empirical data on speech production. The theory-driven approach has also proven instrumental in the development of articulatory speech synthesizers because it reduces the complexity of the vocal tract down to a manageable number of functions.

Here we present a forward model of the articulatory-acoustic mapping that was thoroughly *data-driven* by design. The model was an artificial neural network trained on the articulatory and acoustic recordings from one speaker in the multi-channel articulatory (MOCHA) speech database (Wrench and Hardcastle, 2000), recorded at the Edinburgh speech production recording facility. Inputs to the model were electromagnetic articulograph (EMA), electropalatograph (EPG), and laryngograph (LYG) measurements, each windowed over 64 ms slices of time. The output of the model was a power spectrum of the speech acoustics at the center of each 64 ms slice. The inputs and outputs were coded in the

[a)]Electronic mail: ckello@gmu.edu

model as patterns of activity over sets of connectionist processing units. The mapping from inputs to outputs was governed by a single set of weights on the connections between input and output units, some of which were mediated by hidden units (see Sec. II). Thus, a single, unified set of model parameters had to represent the entire mapping from articulatory inputs to acoustic outputs. The weights were determined by gradient descent learning, which was driven to minimize error between acoustic outputs and their corresponding targets. Acoustic targets were derived from the acoustic recordings.

The model was purely data-driven in that values for the model parameters (i.e., the weights) were learned solely on the basis of articulatory and acoustic data from recorded speech tokens. In other words, the parameters were not *a priori* set on the basis of physical principles of the vocal tract. Moreover, the articulatory and acoustic dimensions were ''raw'' in the sense that they did not directly code articulatory or acoustic features known to convey phonetic information. For instance, articulatory features such as place or manner of articulation were not extracted *a priori* from the articulatory recordings, nor were acoustic features such as formant frequencies. Instead, the articulatory and acoustic data streams were presented to the model in a largely unprocessed format. It is true that some assumptions were built into the model architecture, e.g., that acoustic states could be determined on the basis of a certain window of articulatory data, and that acoustic targets are unimodal (see Sec. II). However, these assumptions were minimal, and in some cases, they were forced by constraints of the articulatory recordings.

The data-driven approach to forward modeling is different from the theory-driven approach in that all empirical data on the vocal tract and the corresponding speech acoustics can be made available to the model. It is the learning procedure and the computational capacity of the model that determines what information is and is not extracted from the data and represented in the model parameters. By contrast, the model parameters in a theory-driven forward model are determined more explicitly by the modeler.

Motivation for a data-driven forward model. In the current modeling work, the data-driven approach was motivated by two aims. First, while theory-driven forward models have proven to be useful research tools, they have not yet enabled the development of natural-sounding articulatory speech synthesizers. One reason for this shortcoming is that, in a theory-driven forward model, many details of the vocal tract and speech acoustics are purposely abstracted away. It is presumably these details (among other factors) that impart the quality of a person's voice. Therefore, one way to achieve more natural-sounding speech synthesis would be to capture as much detail as possible about the vocal tract and speech acoustics for a given speaker (e.g., see also Blackburn and Young, 2000; Jiang *et al.*, 2002; Rowels, 1999; Shiga and King, 2003). The data-driven approach to forward modeling has the potential to capture such details. Relatively little detail about the vocal tract was available for use in the current work (see Sec. III A), but the simulations provided an initial test of the viability of a data-driven articulatory speech synthesizer.

The second aim of the current work was to take a first step toward building a computational model of phonological development. A fundamental question in research on speech acquisition is how does the infant language learner acquire knowledge about the phonological structure of his or her language. Moreover, how is that knowledge represented in the mind and brain of the learner, and then used in language tasks such as spoken word comprehension and production? In recent years, computational models of speech acquisition and production have been developed as tools for exploring and testing the underlying theories (Bailly, 1997; Guenther, 1994, 1995; Guenther, Hampson, and Johnson, 1998; Plaut and Kello, 1999). An integral component of these models is the simulation of babbling and early attempts at the production of spoken words. The forward mapping from articulation to acoustics is essential for such simulations.

The forward model reported here is planned to be one component of a computational model of spoken word acquisition and processing. Plaut and Kello (1999) presented a connectionist model of spoken word acquisition and processing in which distributed representations were learned in the service of speech tasks. The central hypothesis tested in that model was that a learned level of representation exists to (1) integrate the speech signal over word-sized units, (2) generate articulatory trajectories over word-sized units, and (3) map between spoken word forms and their meanings. This level of representation was termed ''phonology'' because its structure was hypothesized to be phonological in nature by virtue of the three core speech tasks that it supported. Thus, the model was aimed at simulating how phonological representations emerge over the course of spoken word acquisition.

On the theoretical approach taken by Plaut and Kello (1999), a central factor in the emergence of phonological representations was their dual purpose in supporting both speech perception and speech production (see Hickok, 2001, for neuroimaging evidence of the existence of dual-purpose representations). As a result of this dual purpose, phonological representations were hypothesized to be shaped, in part, by the intersection of acoustic and articulatory structure in speech. The question, then, is, how does the learning that occurs during the early experiences of speech perception combine with the complementary learning that occurs during speech production to form this intersection. One key part of the answer to this question on the approach taken by Plaut and Kello was that the language learner uses her knowledge of the forward mapping from articulation to acoustics as a bridge between learning in speech perception and learning in speech production. This knowledge was embodied as a forward model of the articulatory-acoustic mapping, and the forward model was learned through simulated babbling (see also Perkell *et al.*, 1997; Perkell *et al.*, 2000).

The forward model played a relatively minor, but absolutely necessary, role in the development of phonological representations. It enabled learning on the input side of the system (i.e., perception and comprehension) to drive learning on the output side of the system. It was not part of the

mechanism that integrated inputs over time to form phonological representations, nor was it part of the mechanism that generated outputs over time to produce articulatory trajectories. Therefore, the current work is intended only to investigate whether the bridge between perception and production can be based on real speech. The forward model is fairly small piece of theory proposed by Plaut and Kello (1999), but the use of real speech would be a major improvement over the original modeling work.

In the Plaut and Kello (1999) simulation, the articulatory and acoustic representations were engineered on the basis of knowledge accumulated over years of phonetics research (e.g., Ladefoged, 1993). For example, articulatory representations included tongue height and backness, and acoustic representations included first through third formant frequencies. A forward mapping from articulatory to acoustic representations was also engineered on the basis of phonetics theory and research, and the task of the forward model was to learn this mapping. Thus, the engineered forward mapping was clearly theory-driven in that it was not derived directly from measurements of speech. As a similar example, Guenther's DIVA model (so named because it maps orosensory Directions Into Velocities of Articulators) of speech acquisition and production also includes a forward model that is based on engineered representations of speech articulations and acoustics (Guenther, 1994, 1995; Guenther *et al.*, 1998).

The simulations reported by Plaut and Kello (1999) and Guenther (1994; Guenther, 1995; Guenther *et al.*, 1998) have been successful in accounting for certain phenomena in speech acquisition and production. For instance, the Plaut and Kello model was able to learn representations that functioned to support the tasks of spoken word comprehension, production, and imitation. The DIVA model has accounted for phenomena of coarticulation, motor equivalence, and speaking rate (among others). Part of what made these successes possible were the simplifying assumptions of the models. Most relevant to the current discussion are the simplifications that were made in the articulatory and acoustic representations, and in the forward mapping between them. These simplifications made the models tractable, and they removed extraneous details that would have made it difficult to relate simulation results to the theoretical principles embodied in the models.

While recognizing the value of theory-driven models, it is fair to ask whether the models offered by Plaut and Kello (1999) and Guenther (1994; Guenther, 1995; Guenther *et al.*, 1998) would scale to handle all of the complexities inherent in the development and processing of real speech. The simulations are meant to serve as evidence for theories of speech acquisition and speech production. However, it is unclear how the models would perform when implemented with more veridical representations of speech articulations and acoustics. If the models were to fail under more veridical conditions, one would have to ask whether the theories were fundamentally flawed in some or way, or whether the failures were only due to shortcomings in the computational machinery.

The use of simplified articulatory and acoustic representations also raises questions about the successes of the models, especially with respect to the Plaut and Kello (1999) model. The most relevant question for the current discussion is the following: does the simulated learning of phonological representations stand as support for Plaut and Kello's theory of phonological development, or did this success depend crucially on the simplifications in the articulatory and acoustic representations? For instance, a major issue in phonological development is how sensitivity to the segmental structure of speech emerges from language experience (e.g., see Bernhardt and Stemberger, 1998; Jusczyk, 1997). On the theory proposed by Plaut and Kello, sensitivity to segmental structure is primarily a product of the articulatory and acoustic structure of speech, and the statistical regularities in the speech inputs that come from adults and other children (of course, neuroanatomy, neurophysiology, and mechanisms of learning also play their respective roles). However, in the simulation reported by Plaut and Kello, segmental structure was partially engineered into the articulatory and acoustic representations. This engineering may have been key to the learning of phonological representations in that simulation. Thus, the simulation results left open the question of whether phonological representations can be learned from articulatory and acoustic representations in which no segmental structure is imposed.

The forward model reported here is a first step toward addressing these and other questions. The articulatory inputs and acoustic outputs used in the current model were derived directly from articulatory and acoustic recordings of a female speaker of British English. The procedures for preprocessing the recordings were designed such that segmental structure was not pre-extracted from the data streams. This is not to say that segmental structure is unimportant to speech; a key test of any model of phonological development would to be show that it is sensitive to the segmental structure of speech in the same way that humans are. The point here is that a complete model would need to explain *how* the language learner becomes sensitive to segmental structure in the native language, given only the raw speech signal as input. The forward models reported here do not explain this aspect of learning; on the Plaut and Kello (1999) theory, the learning of segmental structure is explained by other mechanisms. What the current work provides is a necessary first step toward building a more complete model of phonological development based on real speech.

In addition to this long-term purpose, the current work also served two more immediate purposes. One immediate purpose was to test the viability of a data-driven articulatory speech synthesizer. To the extent that the reported forward models are successful, they will output acoustics that are identical to that of the targets derived from the speaker. However, it is important to note that a forward model does not constitute a speech synthesizer because it does not specify how to control the articulatory dimensions (in the current work, articulatory states always came from the speech database). Nonetheless, a data-driven forward model that can output natural-sounding speech may be an important step toward building a complete, data-driven, articulatory speech synthesizer.

The second immediate purpose of the current work was to generate a lower-bound estimate on the amount of phonetic information that is captured by the articulatory recordings in the MOCHA speech database. The task of the forward models reported herein was to generate the acoustic outputs of articulatory inputs as veridically as possible, defined as the minimization of squared error. If the models could perform this task perfectly, it would mean that the articulatory recordings had captured enough information about the vocal tract to generate all of the acoustic detail in the recordings of the speaker's voice. There was no expectation of perfection, but any phonetic information conveyed by the models had to originate in the articulatory recordings. Therefore, the forward models provided a lower bound on the phonetic information available in the articulatory recordings. The forward models could not provide an upper bound because it is possible that the articulatory recordings contained more phonetic information than measured in the acoustic outputs; there is no guarantee that all phonetic information was extracted by the models, or conveyed by our measures of phonetic information.

II. MODEL

Three forward models were trained on the recordings for one speaker in the MOCHA database. The *all* model was trained on all 460 sentences recorded by the speaker, the *even* model was trained on the even-numbered sentences, and the *odd* model was trained on the odd-numbered sentences. The odd/even split was arbitrary, and was used to test the generalization of the learned model parameters to inputs that were not presented during training. Specifically, the odd-numbered sentences were used to test generalization of the even model, and vice versa for the odd model. Tests of generalization served to ensure that the model parameters captured the general relationship of the vocal tract and the resulting acoustics, rather than individual input/output pairings or some unknown peculiarities in the speech database.

A. Speech database and pre-processing

Speech tokens were drawn from one female speaker of British English (subject ID ''fsew,'' southern dialect) in the MOCHA speech database, recorded at the Edinburgh speech production recording facility. The speech corpus consisted of one token each of 460 phonetically compact sentences designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. The corpus included all 450 phonetically compact TIMIT (sx) sentences, plus ten additional sentences designed to include phonetic pairs and contexts that are particular to British English.

Articulatory recordings in the MOCHA database consists of electromagnetic articulograph (EMA), electropalatograph (EPG), and laryngograph (LYN) recordings. The EMA recordings consisted of eight sensors placed in the midsagittal plane of the vocal tract, attached to the following locations: the vermilion border of the upper lip, the vermilion border of the lower lip, the upper incisor, the lower incisor, the tongue tip (5–10 mm from the tip), the tongue blade (approximately 2–3 cm posterior to the tongue tip sensor), the tongue dorsum (approximately 2–3 cm posterior to the tongue blade sensor), and the soft palate (approximately 10–20 mm from the edge of the hard palate). $[X,Y]$ positions were recorded from each sensor, sampled at 500 Hz.

The positions of these eight sensors were used to calculate nine $[X,Y]$ pairs of articulatory dimensions, as follows. One $[X,Y]$ pair coded the position of the lower incisor (i.e., jaw movement) relative to the upper incisor. This relative coding removed head movement because the position of the upper incisor sensor was fixed relative to the head. Two $[X,Y]$ pairs coded the positions of the upper and lower lips, relative to the positions of the upper and lower incisors, respectively. These pairs coded lip movement independent of head and jaw movement. One $[X,Y]$ pair coded movement of the soft palate relative to the upper incisor. Two $[X,Y]$ pairs coded the overall position of the tongue as the average of the three tongue sensors, one pair in absolute coordinates, and one pair relative to the upper incisor. Finally, three $[X,Y]$ pairs coded each of the three tongue positions, relative to the absolute average tongue position. These three pairs coded local movements of the individual sensors independent of more global movements of the entire tongue.

EPG sensors were placed in 48 normalized positions on the hard palate defined by landmarks on the upper maxilla. Contact between the tongue and each EPG sensor (binary values) was sampled at 200 Hz. LYN recordings provided voicing information at the larynx as a wave form sampled 16 kHz, stored with 16 bit precision, and low-pass filtered at 400 Hz. Acoustic recordings were also sampled at 16 kHz and stored with 16 bit precision, but they were low-pass filtered at 8 kHz instead of 400 Hz.

The acoustic and LYN recordings were transformed from the time domain to the frequency domain with the use of Matlab's fast Fourier transform (FFT) routine. FFTs were calculated over hamming windows 64 ms wide, taken at 32 ms intervals. We explored a range of widths and found 64 ms to produce the most intelligible reconstructed speech signal (see Sec. III). Given the sample rate of 16 kHz, this procedure resulted in 511 frequency bins of log magnitude per window after discarding the dc offset. Phase information in the acoustic signal was discarded in the FFT conversion because the articulatory recordings were not expected to carry phase information (the loss of phase information was partly responsible for the need for relatively wide processing windows). Only the lower 25 bins were used for the LYN recordings because the signal was low-pass filtered at 400 Hz. The rear 24 EPG sensors were discarded because they were not activated in the recordings for the chosen speaker.

For each dimension in the acoustic, EMA, and LYN data streams, the observed values across the entire data set were rank-ordered, and the smallest 100 values were set equal to the 100th smallest value, and the largest 100 values were set equal to the 100th largest value. This procedure normalized very extreme outliers in each dimension of the data streams, thereby restricting their range. The restricted range for each dimension was then normalized to [0,1]. This normalization procedure was not necessary for the EPG data because those dimensions were already normalized in the range [0,1]. Finally, the EMA and EPG data streams were down-sampled to

31.25 Hz, and aligned with the FFT windows calculated over the acoustic and LYN data streams.

B. Articulatory and acoustic representations

Outputs of the forward models were vectors of real numbers in the range [0,1] that represented the acoustic power spectrum at a given 64 ms slice of time. The vectors were 1022 dimensions in size. For each of the 511 FFT bins, one dimension represented the values in the range [0,0.5), and another dimension represented values in the range [0.5,1]. Values outside of a given unit's range were set to zero on that unit. This output format allowed for better resolution in the model's representations, and separate parameters for learning between the upper and lower ranges (i.e., separate sets of connection weights fed into the upper-range and lower-range output units; see Sec. III C).

Inputs to the forward model were vectors of real numbers in the range [0,1] that represented the previous (32 ms in the past), current, and next (32 ms in the future) articulatory states, relative to the acoustic outputs. The input vectors were 588 dimensions in size, with one third each representing the previous, current, and next articulatory states. Each point in time consisted of 72 dimensions dedicated to EMA positions, 24 dimensions dedicated to EPG contact, and 100 dimensions dedicated to FFT values from the LYN recordings. The EPG dimensions directly coded the average amount of tongue contact in a given slice of time for each of the front 24 EPG sensors. Four dimensions were assigned to each of the 18 EMA dimensions (i.e., nine pairs of $[X,Y]$ positions), and each of the 25 bins of FFT magnitude (up to 400 Hz in frequency) for the LYN recordings. For each quadruple assigned to value x, one dimension coded x directly, one coded the value $1-x$, one coded the x values in the lower range [0,0.5), and one coded x values in the upper range [0.5,1]. Analogous to the output format, the split range format provided the model with a separate set of parameters for the lower and upper ranges of input values. To complement, the $x|1-x$ inverse coding provided two sets of parameters that spanned the full range of input values. The inverse coding was used to ensure that learning occurred on every training example, for each dimension, regardless of each dimension's value. In backpropagation, no learning will occur on a unit's sending weights when the activation value of that unit is zero. Thus, the x and $1-x$ units served to provide model parameters learned on either side of each dimension.

C. Forward model training and results

All three forward models had the neural network architecture depicted in Fig. 1. Each model consisted of 588 input units, 1022 output units, and 100 hidden units. Acoustic representations corresponded to patterns of activity over the output units, and articulatory representations corresponded to patterns of activity over the input units. Patterns of activity over the hidden units corresponded to internal representations that were learned over the course of training (see the following). The activation value for each hidden unit and each output unit was calculated as the sigmoid of the dot product of the unit's incoming weights and the outputs of the pre-synaptic units. The activation values of the input units were set directly equal to the articulatory representation at a given point in time in one of the trained sentence tokens (i.e., composed of previous, current, and next articulatory states as described earlier). Every articulatory input unit was connected to every acoustic output unit (i.e., full connectivity). In addition, articulatory inputs were fully connected to the hidden units, and the hidden units were fully connected to the output units. Direct connections between inputs and outputs were included to increase the rate of learning by facilitating the extraction of any linear relationships between the articulatory and acoustic dimensions. The hidden units served to capture nonlinear relationships between the input and output dimensions, although they were free to capture linear relationships as well. A total of 100 hidden units was chosen on the basis of trial and error; pilot work indicated that model performance was worse with fewer hidden units, and no better with more hidden units.

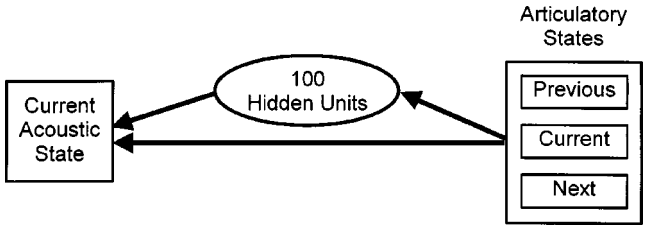

FIG. 1. Architecture of the forward models. Arrows indicate full connectivity between groups of processing units.

At the start of training, the weights on all connections in the network were drawn randomly with replacement from a rectangular distribution in the range $(-0.1,0.1)$. Weights were learned via the backpropagation of error signals generated on the outputs units (Rumelhart, Hinton, and Williams, 1986). In particular, time slices from the sentence tokens were presented to the network in batches of 100, sampled at random from the training set. For each time slice, the activation values of the input units were set to the corresponding articulatory representation, and those activation values were propagated forward through the network connections to generate a pattern of activation on the output units. For each output unit, squared error was calculated between the unit's activation value, and its target activation, which was determined by the acoustic representation for the time slice in question. The error signals were then backpropagated along the network connections to calculate weight derivatives. Each weight's derivatives were summed across each batch of 100 training examples (i.e., time slices). After each batch, the summed derivatives were used to update each weight according to

$$\Delta w_{ij}^{[b]} = \eta_N \eta_{ij} \frac{\partial E}{\partial w_{ij}} + \alpha(\Delta w_{ij}^{[b-1]}), \tag{1}$$

where η_N was the overall network learning rate (decreased from $5e-4$ to $5e-5$ over the course of training), η_{ij} was a weight-specific learning rate, α was a momentum term (fixed at 0.8), and b was the Nth batch over the course of training. Weight-specific learning rates were adjusted on the basis of

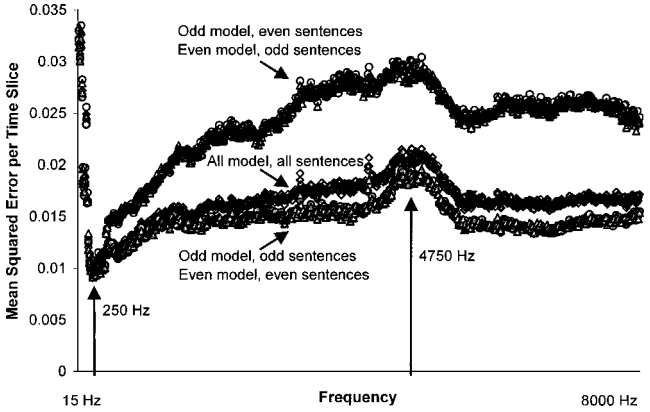

FIG. 2. Mean squared error per output unit per time slice, as a function of frequency for the *all* model, the *even* model, and the *odd* model.

the consistency of weight derivatives across batches (Jacobs, 1988).

The 460 sentences in the training set for the *all* model contained a total of 41 791 slices of time to be trained. The *even* model contained 20 794 slices, and the *odd* model contained 20 997 slices. Each model was trained on 100 000 batches of training examples, which is about the point at which the reduction in error became miniscule. All three models were stopped at exactly 100 000 batches to control for amount of training. The training sets did not appear to be overfit because, for the odd and even models, error on the untrained sentences decreased throughout training. At the end of training, the average squared error between the targets and outputs was calculated per frequency per time slice. These averages are shown in Fig. 2 for each of the three forward models, separated by sentence type (odd- or even-numbered). As can be seen, the mean squared error never exceeded 0.035 by the end of training. By comparison, mean squared error at the beginning of training was 0.193.

Figure 2 shows that there was a clear rank ordering in the overall amount of model error. The models trained on half of the sentences in the speech database (*even* and *odd* models) produced the least amount of error on their respective training sets, and the most amount of error on sentences outside their training sets. The *all* model produced slightly more error than the *even* and *odd* models for their respective training sets, but substantially less error than those models for sentences outside their training sets. This rank-ordering of error indicates that some learning did not generalize beyond the training sets in the *odd* and *even* models, and that error from these models was reduced somewhat by learning that was specialized to the training sets.

The pattern of error as a function of frequency was mostly similar across the different models and training sets. There was a sharp dip in error at about 250 Hz, followed by a fairly steady climb in error to a peak at about 4750 Hz. Error then dropped to a middling baseline level that was maintained out to 8000 Hz. This pattern was somewhat different for the *odd* and *even* models tested outside their training sets in that, for those models, an extra plateau of error can be seen prior to the peak at about 4750 Hz. The dip at 250 Hz is due to the fact that the LYN recordings contain fairly direct information about acoustic energy around the pitch of the speaker's voice. The peak at about 4750 Hz might have been due to the models' inability to determine the spectral details of acoustic energy generated by fricative and plosive speech sounds (but this conjecture needs further investigation). The reasons for the particular characteristics of the rise in error up to its peak, and its drop off after the peak, are currently unknown.

The error scores gave a detailed picture of which frequency bands were processed more or less accurately by the models, but these scores do not give an interpretable measure of intelligibility of the target and model tokens. To provide a more standard measure, an energy histogram method (Hirsch, 1995) was used to estimate the signal-to-noise ratio (SNR) in the target and model tokens, as well as in the original, unprocessed tokens. The mean SNR for each category of tokens was as follows: 28.2 dB for the original tokens, 24.4 dB for the target tokens, 25.8 dB for the *all* model tokens, 26.0 for the *even* model tokens, and 25.7 dB for the *odd* model tokens. These means show that, although the original tokens were distinguished from the processed ones, the energy histogram method did not reflect the overall differences in error scores plotted in Fig. 2. A SNR measure that uses the target signal as a baseline would be more appropriate, but the removal of phase information in the target and model tokens prohibited such a measure. To allow other researchers to experiment with various measures of intelligibility, the wave forms all of the stimuli in all four conditions can be downloaded at http://archlab.gmu.edu/~ckello/forward-models.html.

The error scores and SNRs are quantitative, objective measures of intelligibility. A more qualitative way to assess the modeling results is to view spectrograms of the target utterances, and compare them against spectrograms of the corresponding model outputs. In Figs. 3 and 4, spectrograms are shown for one odd-numbered and one even-numbered example sentence, each chosen arbitrarily from the speech database. At a glance, the model spectrograms are quite similar to the target spectrograms. Some of the spectral and temporal details in the targets appear to be washed out in the model outputs, particularly above 3000 Hz where the harmonics appear to be completely washed out. These spectrograms are informative visualizations, but ultimately, the forward models must be assessed by measuring the amount of phonetic information contained in their outputs. Such an assessment is reported in Sec. III.

III. INTELLIGIBILITY TESTS

The error results shown in Fig. 2 provide a quantitative measure of performance for the forward models, and Figs. 3 and 4 provide a more qualitative measure. However, it is difficult to interpret these measures in terms of the amount of phonetic information that was captured in the forward mapping learned by the models. To better estimate the phonetic information captured in the models, the model outputs and targets were submitted to empirical tests of intelligibility. Intelligibility of the targets served as a baseline comparison. The percentage of words identified correctly was used as a coarse measure of the overall amount of phonetic information captured by the models, relative to the phonetic infor-

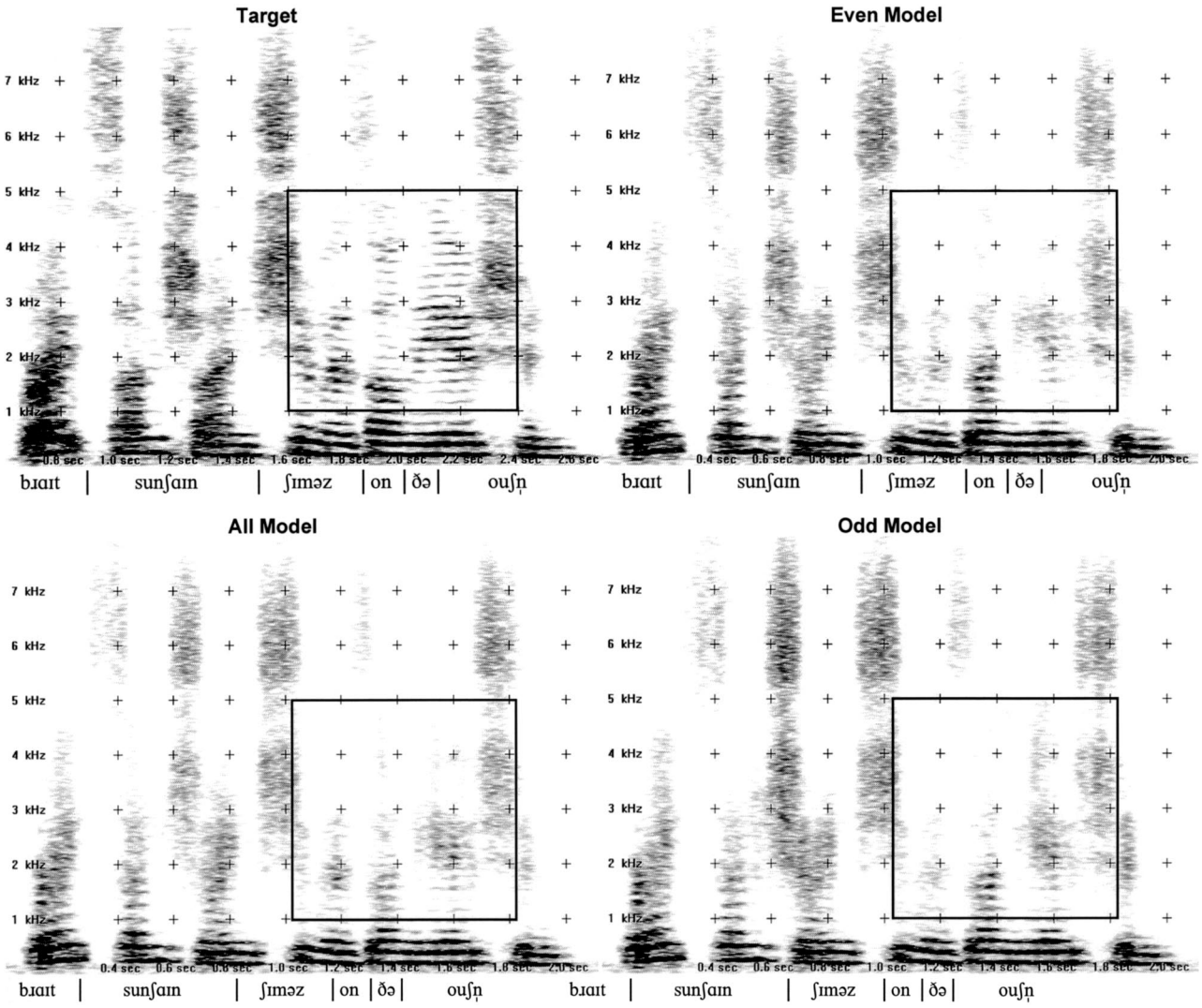

FIG. 3. Spectrograms of the target acoustics, and the acoustics output by each of the three model types for an even-numbered sentence token, ''bright sunshine shimmers on the ocean.'' The outlined region shows where the models can be seen to have lost some of the spectral detail in the target utterance.

mation available in the targets. To provide a rough measure of the kinds of phonetic information that was lost by the models, phoneme confusions were identified in the responses (when possible), and tabulated.

A. Methods

1. Participants

Eight undergraduates participated as listeners in the speech intelligibility tests for course credit. All participants reported being native speakers of American English, none reported a hearing impairment, and none were familiar with the TIMIT speech database.

2. Stimuli

All 460 sentence tokens were passed through each of the three forward models to generate a series of acoustic outputs for each token, and from each model. The Matlab inverse FFT routine was used to convert the acoustic outputs into an acoustic wave form for each sentence token, from each model. The same procedure was also applied to the targets, resulting in four stimulus tokens for each sentence: one from each of the three model types, and one from the target. As noted earlier, some information in the original acoustic recordings was lost because it was necessary to discard phase information in the FFT procedure used to generate the acoustic targets for the models. Phase information was replaced by inserting random phases into the inverse FFT procedure. Pilot tests indicated that random phases produced more intelligible wave forms compared with phases fixed at values such as zero. However, loss of the original phase information caused some distortion in the generated wave forms.

3. Procedure

Participants were seated in a quiet booth and instructed that they would be listening to grammatically correct and semantically plausible English sentences. For each sentence, they were instructed to transcribe what they heard to the best of their ability. They were told that some sentences were garbled and therefore difficult to hear. They were asked to type into the computer as many words as they heard for each

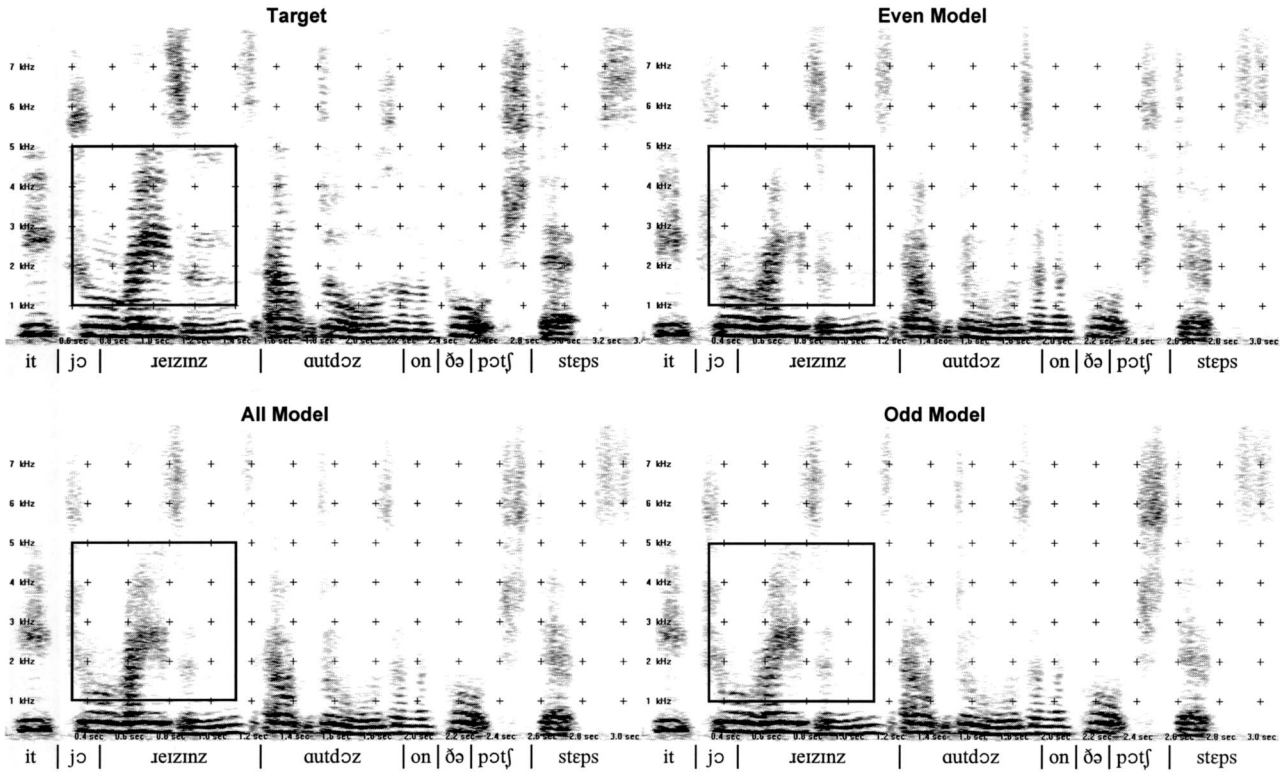

FIG. 4. Spectrograms of the target acoustics, and the acoustics output by each of the three model types for an odd-numbered sentence token, ''eat your raisins outdoors on the porch steps.'' The outlined region shows where the models can be seen to have lost some of the spectral detail in the target utterance.

sentence, in the order that they heard them, and they were encouraged to guess at words whenever necessary and possible.

Stimulus presentation and data collection was controlled through a graphical user interface, and stimuli were presented over Sennheiser MH80 headphones at a comfortable listening level that was held constant across participants. Each trial began with the participant clicking on a button to listen to the current sentence. Participants were forced to click on this button three times in order to listen to each sentence three times before responding. Participants typed each response into a text entry field, and clicked on another button to enter the response and begin the next trial. Participants were not given feedback at any time.

Participants were given four practice sentences at the beginning of the experiment, followed by one-fourth (115) of the 460 sentence tokens. Tokens were rotated across subjects to cover all 460 sentences evenly, and each sentence appeared in two of the four token conditions. The token conditions were rotated across subjects such that each condition was sampled an equal number of times.

B. Results

All responses were corrected for spelling errors, and in the few cases where the participant responded with a homophone of the correct word response (e.g., responding with TACKS when the correct word is TAX), the homophone was replaced with the correct word. The percentage of words transcribed correctly for the even-numbered and odd-numbered sentences is graphed in Fig. 5 for each of the four token conditions. The graph shows that the target outputs were transcribed most accurately, with no noticeable difference in accuracies for the even-numbered versus odd-numbered sentences. There are at least three possible reasons why the intelligibility of the targets was less than perfect: (1) the tokens were generated by a speaker of British English, but the listeners were speakers of American English, (2) some of the TIMIT sentences contain words likely to be unfamiliar to the participants (e.g., ''neoclassic,'' ''Nan,'' ''statuesque,'' etc.), and (3) phase information in the original recordings was lost in the FFT procedure.

The graph also shows that accuracy for the outputs of the *all* model was 11 percentage points lower on average than that for the targets, $t(7)=7.4$, $p<0.001$. Compared with the *all* model, similar levels of accuracy were found for

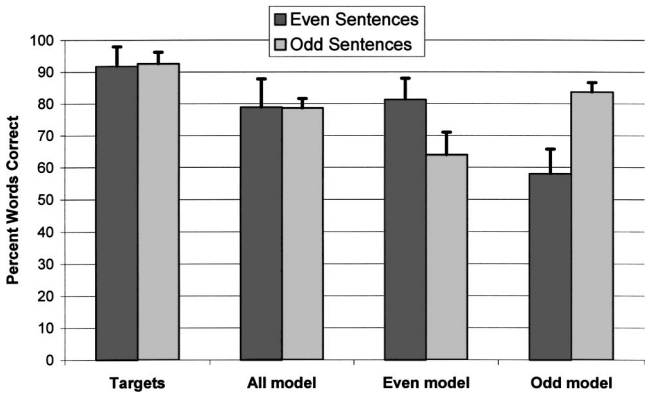

FIG. 5. Mean percent words correct in the intelligibility tests, as a function of token type (target, *all* model, *even* model, or *odd* model) and sentence type (even-numbered or odd-numbered sentences). Error bars show standard deviations of the subject means.

TABLE I. Counts of phoneme confusions summed across all model conditions and all subjects, with the number of times that each confused phoneme appeared in the corpus included as a baseline. Counts were collapsed across order of confusion, and counts under three were not included.

Confusion	Articulatory dimensions confused	No. Observed	No. in corpus
/l, ɹ/	Place	9	1279
/p, b/	Voicing	6	677
/m, b/	Manner	4	717
/n, d/	Manner, nasal	4	1343
/n, l/	Nasal, lateral	4	1361
/k, p	Place	4	901
/n, m/	Place	4	1301
/j, d/	Place, manner	4	669
/k, h/	Place, manner	4	706
/k, f/	Place, manner	4	797
/j, t/	Place, manner, voicing	4	1008
/n, r/	Place, nasal	4	1548
/k, b/	Place, voicing	4	846
/k, t/	Place	3	1402
/b, f/	Place, manner, voicing	3	573
/ç, j/	Voicing	3	253
/d, t/	Voicing	3	1395

the *even* model tested on the even-numbered sentences and the *odd* model tested on the odd-numbered sentences. These model results are reported primarily as points of comparison for the tests of generalization. In particular, accuracy was 20 percentage points lower for the *even* model tested on the odd-numbered sentences, compared with the same model tested on the even-numbered sentences, $t(6)=3.5$, $p<0.05$. Similarly, accuracy was 18 percentage points lower for the *odd* model tested on the even-numbered sentences, compared with the same model tested on the odd-numbered sentences, $t(6)=6.1$, $p<0.001$.

To provide a rough measure of the kinds of phonetic information that were lost by the models, responses to all model outputs were inspected for phoneme confusions. Responses were identified in which a target word in the sentence was clearly replaced with a different word in the response. On this strict criterion, replacements were identified for only 57% of the responses with errors. The difficulty was that participants often left target words out of their responses, or occasionally inserted words that were not in the target sentences. Deletions and insertions often made it difficult to align a given response with its target. To avoid experimenter bias in alignment decisions, no word replacements were identified when the alignment was ambiguous.

The counts of phoneme confusions are shown in Table I. These counts are collapsed across confusion order (i.e., phoneme A replaced with phoneme B, or B with A), and only counts greater than two are shown. In the full set of confusions, vowels were never confused with consonants, and of the few vowel–vowel confusions that were identified, no particular vowel–vowel confusion occurred more than twice. With respect to consonants, the phoneme pairs /l, ɹ/ and /p, b/ were confused most often. This was true in terms of raw counts, and counts relative to the number of times these phonemes appeared in the corpus. Otherwise, features that denote place of articulation were confused most often, but features denoting voicing and manner were also confused with some regularity. Confused features were often similar to each on their respective dimension of confusion; for example, the feature plosive was often confused with the feature affricate, and both manners of articulation are characterized by a burst release. Beyond these general statements, it is difficult to be more specific without results from an experiment aimed more directly at phoneme identification. The MOCHA database contained only sentence stimuli, which made it prohibitively difficult to conduct a phoneme identification experiment.

IV. DISCUSSION

Three neural network models were trained on the articulatory-acoustic mapping for one speaker in the MOCHA speech database. Results indicated that this mapping was well-approximated in the models. Spectrograms and analyses of model error showed that the acoustic outputs in lower frequency range (below 2000 Hz) closely matched the target outputs, whereas acoustic outputs in the upper frequency range (above 3000 Hz) were less accurate. Intelligibility tests showed that listeners could identify a large percentage of words in sentences that were generated by passing the recorded articulatory trajectories through the models. These tests also showed that the model parameters generalized, to some degree, to novel articulatory inputs. On the one hand, intelligibility of the untrained sentences (61% words correct on average) demonstrated that the model learned something about the general relationship between articulatory and acoustic parameters for the speaker's vocal tract. On the other hand, reduced intelligibility of the untrained sentences compared with the trained sentences (81% words correct on average) indicated that some aspects of the general articulatory-acoustic relationship were not learned sufficiently.

The intelligibility tests provided coarse measures of phonetic information in that phoneme confusions provided only a rough measure of the kinds of phonetic information that were and were not contained in the acoustic outputs. Other measures of phonetic information, such as those derived from tests of phoneme identification (e.g., Bernstein, Demorest, and Tucker, 2000), would provide more detail about phonetic information in the model outputs. Unfortunately, only the sentence recordings were available for intelligibility tests, and it would have been difficult to specifically test phoneme identification with sentence stimuli.

Nonetheless, the results in hand demonstrate that the forward mapping from articulations to acoustics can be learned, at least to a reasonable extent, via a heuristic of gradient descent (i.e., backpropagation) in an acoustic error space. They also place a lower bound on the amount of phonetic information captured by the articulatory recordings in the MOCHA database. In particular, articulatory recordings were comprised of 8 mid-sagittal $[X,Y]$ positions at key locations in the vocal tract, 24 positions of tongue contact with the hard palate, and FFTs of the acoustic energy generated at the larynx. These articulatory recordings were sufficient to generate much of the spectral and temporal information in the resulting speech acoustics. The phonetic information in the

models' outputs had to come from the articulatory inputs because the models were data-driven, i.e., they did not contain any *a priori* information about speech.

In fact, it is possible that the articulatory recordings actually contained more phonetic information than indicated by the reported models. Some information was lost in preprocessing the articulatory recordings in order to format them for the models (e.g., phase information was lost in the FFT procedure), and some of this lost information may have been phonetic in nature. Even if no information was lost in pre-processing, it is possible that the mapping defined by the model parameters (connection weights) did not capture all of the phonetic information in the articulatory inputs. This shortcoming can occur because of an inadequacy in back-propagation, in the use of sigmoidal processing units, or in the representational scheme used on the inputs or outputs. Gradient descent learning can settle into a local minimum in error. While any differentiable function can be approximated using sigmoidal hidden units (Cybenko, 1989), some functions are better suited than others for this particular basis function, and the generalization of learning can be influenced by the choice of activation function (Rumelhart *et al.*, 1995). Finally, it is well known that the design of input and output representations is critical to learning and performance in all neural networks (e.g., Plaut *et al.*, 1996). Thus, it is possible that an alternate method of modeling would have resulted in a forward mapping that captured more phonetic information than the models reported herein.

One reason to improve the fidelity of the current forward models is for the purpose of an articulatory speech synthesizer. Acoustic outputs of the reported models were natural-sounding in that they captured the quality of the speaker's voice, although limitations of the models caused their outputs to sound as if they were masked by noise of some kind. Thus, the models might contribute to the development of a natural-sounding articulatory synthesizer if their fidelity was improved. However, a formidable hurdle in such an effort would be to manipulate the articulatory dimensions such that any desired utterance could be produced. All sequences of model outputs reported in the current work were generated from articulatory sequences in the speech database. However, a speech synthesizer must be able to synthesize any given sequence of phones. Modeling articulatory trajectories is known to be a difficult problem (e.g., Kaburagi and Honda, 2001), and the large number of articulatory dimensions used in the current models are likely to exacerbate this problem. Traditionally, articulatory degrees of freedom are reduced and made independent by means of theoretical (Mermelstein, 1973) or empirical (e.g., Badin *et al.*, 2002; Beautemps, Badin, and Bailly, 2001; Blackburn and Young, 2000b) methods. Such methods could be applied to the current forward models, or alternatively, a concatenative method (see Chappell and Hansen, 2002) could be applied to articulatory trajectories recorded specifically for phones or diphones. In any case, further work is necessary to determine whether the forward models reported here could be used in an articulatory speech synthesizer.

Just as a forward model is only one possible component of an articulatory speech synthesizer, it is also only one possible component of a full-scale model of speech acquisition and production. As argued in Sec. I, it would be informative to test whether models of speech acquisition and production can handle the complexities of real speech. The incorporation of a data-driven forward model, similar to the models reported here, would be a significant step toward such a test. However, some difficult problems would need to be addressed before a complete model could be implemented.

For instance, the problem of articulatory control that confronts the development of articulatory speech synthesizers would also confront the development of models of speech acquisition and production. Guenther's DIVA model (Guenther, 1994, 1995; Guenther *et al.*, 1998) has accounted for a number of phenomena that are relevant to the issue of articulatory control, but it is currently unknown whether the DIVA model would scale to handle the control of a real human vocal tract. The Plaut and Kello (1999) approach is well-suited to forward models such as the ones reported here, given that both share the same mechanisms of neural network learning and processing. However, it is currently unknown whether such mechanisms are capable of learning phonological representations on the basis of real speech input. Another issue that would have to be confronted is variability in the speech signal (e.g., see Perkell and Klatt, 1986; Pisoni, 1981). For instance, on any given occasion, the speech signal that corresponds to a given word will be shaped by factors such as the linguistic and nonlinguistic context, and the talker's dialect and voice quality. The resultant variability poses a significant challenge for any effort to build a computational model of speech acquisition and production. The modeling work reported here is one step toward meeting these and other challenges inherent to the research and engineering of speech.

ACKNOWLEDGMENTS

We would like to thank Brandon Beltz, Laura Leach, and Dana Morgan for conducting the intelligibility tests. We would like to also thank Dana Morgan for coding the data from the intelligibility tests. This work was funded in part by NIMH Grant No. MH55628, and NSF Grant No. 0239595.

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., and Savariaux, C. (**2002**). "Three-dimensional linear articulatory modeling of tongue, lips, and face, based on MRI and video images," J. Phonetics **30**, 533–553.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). "Analysis of vocal-tract shape and dimensions using magnetic-resonance-imaging-vowels," J. Acoust. Soc. Am. **90**, 799–828.

Bailly, G. (**1997**). "Learning to speak. Sensori-motor control of speech movements," Speech Commun. **22**, 251–267.

Beautemps, D., Badin, P., and Bailly, G. (**2001**). "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," J. Acoust. Soc. Am. **109**, 2165–2180.

Beautemps, D., Badin, P., and Laboissiere, R. (**1995**). "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies—A new model for vowels and fricative consonants based on experimental-data," Speech Commun. **16**, 27–47.

Bernhardt, B. H., and Stemberger, J. P. (**1998**). *Handbook of Phonological Development: From the Perspective of Constraint-based Nonlinear Phonology* (Academic, San Diego).

Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (**2000**). "Speech perception without hearing," Percept. Psychophys. **62**, 233–252.

Blackburn, C. S., and Young, S. J. (**2000**). ''A self-learning predictive model of articulator movements during speech production,'' J. Acoust. Soc. Am. **107**, 1659–1670.
Blackburn, C. S., and Young, S. (**2000a**). ''Enhanced speech recognition using an articulatory production model trained on X-ray data,'' Comput. Speech Lang. **15**, 195–215.
Blackburn, C. S., and Young, S. (**2000b**). ''A self-learning predictive model of articulator movements during speech production,'' J. Acoust. Soc. Am. **107**, 1659–1670.
Chappell, D. T., and Hansen, J. H. L. (**2002**). ''A comparison of spectral smoothing methods for segment concatenation based speech synthesis,'' Speech Commun. **36**, 343–374.
Cybenko, G. (**1989**). ''Approximation by superpositions of a sigmoid function,'' Math. Control, Signals, Syst. **2**, 303–314.
Goodyear, C. C. (**2000**). ''Incorporating lip protrusion and larynx lowering into a time domain model for articulatory speech synthesis,'' Comput. Speech Lang. **14**, 211–226.
Greenwood, A. R., Goodyear, C. C., and Martin, P. A. (**1992**). ''Measurements of vocal-tract shapes using magnetic-resonance-imaging,'' IEEE Proc.-I: Commun. Speech Vision **139**, 553–560.
Guenther, F. H. (**1994**). ''A neural-network model of speech acquisition and motor equivalent speech production,'' Biol. Cybern. **72**, 43–53.
Guenther, F. H. (**1995**). ''Speech sound acquisition, coarticulation, and rate effects in a neural-network model of speech production,'' Psychol. Rev. **102**, 594–621.
Guenther, F. H., Hampson, M., and Johnson, D. (**1998**). ''A theoretical investigation of reference frames for the planning of speech movements,'' Psychol. Rev. **105**, 611–633.
Hickok, G. (**2001**). ''Functional anatomy of speech perception and speech production: Psycholinguistic implications,'' J. Psycholinguist. Res. **30**, 225–235.
Hirsch, H. G., and Ehrlicher, C. (**1995**). ''Noise estimation techniques for robust speech recognition,'' Paper presented at the Proc. ICASSP.
Jacobs, R. A. (**1988**). ''Increased rates of convergence through learning rate adaptation,'' Neural Networks **1**, 295–307.
Jiang, J. T., Alwan, A., Keating, P. A., Auer, E. T., and Bernstein, L. E. (**2002**). ''On the relationship between face movements, tongue movements, and speech acoustics,'' Eurasip J. Appl. Signal Process. **2002**, 1174–1188.
Jordan, M. I., and Rumelhart, D. E. (**1992**). ''Forward models—Supervised learning with a distal teacher,'' Cogn. Sci. **16**, 307–354.
Jusczyk, P. W. (**1997**). *The Discovery of Spoken Language* (MIT, Cambridge).
Kaburagi, T., and Honda, M. (**2001**). ''Dynamic articulatory model based on multidimensional invariant-feature task representation,'' J. Acoust. Soc. Am. **110**, 441–452.
Kelly, J. L., and Lochbaum, C. C. (**1962**). ''Speech Synthesis,'' Proceedings of the Fourth International Congress of Acoustics, paper G42, 1–4, in *Speech Synthesis*, edited by J. L. Flanagan and L. R. Rabiner (Dowden, Hutchinson & Ross, Stroudsburg, PA), pp. 127–130.
Ladefoged, P. (**1993**). *A Course in Phonetics* (Harcourt Brace, Orlando, FL).
Mermelstein, P. (**1973**). ''Articulatory model for the study of speech production,'' J. Acoust. Soc. Am. **53**, 1070–1082.
Perkell, J., and Klatt, D. (Eds.). (**1986**). *Invariance and Variability in Speech Processes*. (Lawrence Erlbaum, Hillsdale, NJ).
Perkell, J. Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., *et al.* (**1997**). ''Speech motor control: Acoustic goals, saturation effects, auditory feedback, and internal models,'' Speech Commun. **22**, 227–250.
Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J. *et al.* (**2000**). ''A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss,'' J. Phonetics **28**, 233–272.
Pisoni, D. B. (**1981**). ''Some current theoretical issues in speech perception,'' Cognition **10**, 249–259.
Plaut, D. C., and Kello, C. T. (**1999**). ''The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach,'' in *The Emergence of Language*, edited by B. MacWhinney (Erlbaum, Mahweh, NJ), pp. 381–415.
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (**1996**). ''Understanding normal and impaired word reading: Computational principles in quasi-regular domains,'' Psychol. Rev. **103**, 56–115.
Roweis, S. (**1999**). ''Data driven production models for speech processing,'' Ph.D. thesis, University of Toronto, Toronto.
Rubin, P., Baer, T., and Mermelstein, P. (**1981**). ''An articulatory synthesizer for perceptual research,'' J. Acoust. Soc. Am. **70**, 321–328.
Rumelhart, D. E., Durbin, R., Golden, R., and Chauvin, Y. (**1995**). ''Backpropagation: The basic theory,'' in *Backpropagation: Theory, Architectures, and Applications Developments in Connectionist Theory*, edited by Y. Chauvin and D. E. Rumelhart (Erlbaum, Hillsdale, NJ), pp. 1–34.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (**1986**). ''Learning representations by back-propagating errors,'' Nature (London) **323**, 533–536.
Shiga, Y., and King, S. (**2003**). ''Estimation of voice source and vocal tract characteristics based on multi-frame analysis.'' Paper presented at Eurospeech.
Wrench, A., and Hardcastle, W. (**2000**). ''A multichannel articulatory speech database and its application for automatic speech recognition,'' Paper presented at the Proceedings of the 5th Seminar on Speech Production.