Statistical Methods for Estimation of Direct and Differential Kinematics of the Vocal Tract

Adam Lammert[a,*], Louis Goldstein[b,c], Shrikanth Narayanan[a,b], Khalil Iskarous[b,c]

[a] *Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Ave., Los Angeles, CA 90089, USA*[☆]
[b] *Department of Linguistics, University of Southern California, Grace Ford Salvatory 301, Los Angeles, CA 90089-1693, USA*
[c] *Haskins Laboratories, 300 George Street, Suite 900, New Haven, CT, 06511, USA*

Abstract

We present and evaluate two statistical methods for estimating kinematic relationships of the speech production system: Artificial Neural Networks and Locally-Weighted Regression. The work is motivated by the need to characterize this motor system, with particular focus on estimating differential aspects of kinematics. Kinematic analysis will facilitate progress in a variety of areas, including the nature of speech production goals, articulatory redundancy and, relatedly, acoustic-to-articulatory inversion. Statistical methods must be used to estimate these relationships from data since they are infeasible to express in closed form. Statistical models are optimized and evaluated – using a heldout data validation procedure – on two sets of synthetic speech data. The theoretical and practical advantages of both methods are also discussed. It is shown that both direct and differential kinematics can be estimated with high accuracy, even for complex, nonlinear relationships. Locally-Weighted Regression displays the best overall performance, which may be due to practical advantages in its training procedure. Moreover, accurate estimation can be achieved using only a modest amount of training data, as judged by convergence of performance. The algorithms are also applied to real-time MRI data, and the results are generally consistent with those obtained from synthetic data.

Keywords: speech production, direct kinematics, differential kinematics, task dynamics, articulatory synthesis, kinematic estimation, statistical machine learning, locally-weighted regression, artificial neural networks

1. Introduction

The kinematics of complex motor systems can be described at different levels of abstraction (Bernstein, 1967; Hollerbach, 1982; Saltzman and Kelso, 1987). One classic example is the arm – whether human or robotic – which can variously be described as a collection joint angles or by the spatial coordinates of the end-effector. Similarly, the speech production system can be described at several levels, including muscle activations, constriction degrees/locations or formant frequencies. Thus, the choice of variables for describing a system can be low-level (i.e., close to the articulatory substrate) or high-level (i.e., removed from the articulators). This multi-level view of motor systems is common to many studies of biological motor activity (Soechting, 1982; Mottet et al., 2001), and has also been extensively studied in robotic control (Khatib, 1987; Nakanishi et al., 2008). In order to completely characterize a motor system, it is crucial to understand the maps that relate these different kinematic levels. A fundamental understanding of the speech production system can also be built on an understanding these relationships.

The importance of these maps in characterizing motor systems is underscored by the fact that the goals of movement are often defined in terms of relatively higher-level variables, rather than at the level of the articulatory substrate. The variables which are used to define these goals can be called *task variables*, and the space defined by those variables is known as *task space*. Similarly, the lowest level of description is defined by *articulator variables* in *articulator space*. It is often convenient and desirable for a system to be controlled in task space. There is strong empirical evidence supporting the idea that control of speech production is done at a higher level than muscle activations. For instance, it has been shown that the low-level articulators are kept compliant during speech production, allowing the achievement of higher-level tasks even despite per-

☆The SAIL homepage is `http://sail.usc.edu`

*Corresponding author. Tel.: +1 919 724 5364; fax: +1 213 740 4651.
Email address: `lammert@usc.edu` (Adam Lammert)

turbation (Abbs and Gracco, 1984; Kelso et al., 1984; Guigon et al., 2007). Knowledge of the map between articulator space and task space is a prerequisite for being able to accomplish this.

A significant challenge for system characterization, then, is posed by the fact that these kinematic relationships are complex, nonlinear and infeasible to express in closed form for complex motor systems (Sciavicco and Siciliano, 2005). This is certainly true of the speech production system. For instance, if one considers articulatory variables to be individual muscle activations and the task variables to be more abstract quantities like constriction degrees or formant frequencies, then the map will represent a variety of physical processes between those levels. There have been, of course, painstaking efforts to develop such models for both research and speech synthesis, built upon knowledge of vocal tract geometry (Rubin et al., 1996; Iskarous et al., 2003; Nam et al., 2004, 2006) and biomechanical knowledge (Perrier et al., 1996; Payan and Perrier, 1997; Perrier et al., 2003; Gérard et al., 2003; Fels et al., 2005; Vogt et al., 2005, 2006; Gérard et al., 2006; Winkler et al., 2011b,a). Fortunately, it is possible to directly and statistically estimate the maps from data when they cannot be expressed succinctly. Building models in this way should be especially useful and timely in light of the accelerating availability of rich, complete kinematic data of speech (e.g., Wrench and Hardcastle (2000); Narayanan et al. (2004, 2011)).

Attempts have been made to statistically estimate the *direct kinematics*, which express task variables as a function of articulator variables. There has long been a need to statistically examine the direct kinematics of robotic systems for the purposes of calibration. By admitting that any idealized mathematical description of a motor system will differ from its actual physical instantiation, there arises a need to tune the parameters of the system to ensure proper control. Traditionally, the functional form of the relationship is known *a priori*, the problem consists of estimating only the parameters of that form. This can be done statistically, using data from the system in operation. This kind of kinematic calibration is common practice in experimental robotics, and of critical importance in industrial robotic situations (Sklar, 1989; Mooring et al., 1991; Bennet and Hollerbach, 1991; Hollerbach and Wampler, 1996).

Statistical methods have also been employed to estimate the entire functional form of the forward map for a variety of motor systems. Artificial Neural Networks (ANNs) have been used to learn the direct kinematic relationships in the context of articulated robotic arms (Jordan, 1992; Jordan and Rumelhart, 1992) and simulated biological arms (Bullock et al., 1993). Locally-linear techniques have also been employed for estimating robotic forward models (D'Souza et al., 2001; Ting et al., 2008).

Estimation of direct kinematics has also been demonstrated in the domain of speech production. Purely codebook-driven techniques have been utilized for this purpose (Kaburagi and Honda, 1998). Clustering techniques (Shiga and King, 2004) including, perhaps most notably, Gaussian Mixture Models have been used to learn the forward map to a high degree of accuracy (Toda et al., 2004, 2008). ANNs have also been used to estimate the direct kinematics of the speech production system (Bailly et al., 1991; Kello and Plaut, 2004), most prominently as part of developing the DIVA model (Guenther, 1994, 1995; Guenther et al., 1998). Hidden Markov Models have also been used, usually for applications in speech synthesis (Hiroya and Honda, 2002a,b, 2003, 2004; Nakamura et al., 2006). Locally-linear techniques have also been recently used for estimating the map between fleshpoints on the tongue and the formant frequencies of vowels (McGowan and Berger, 2009).

This work is aimed at identifying reliable algorithms that hold promise for estimating the direct and, crucially, the differential kinematics of speech production from speech articulation data. *Differential kinematics* relate velocities in task space with velocities in articulator space (i.e., the first-order partial derivatives of task variables with respect to articulator variables). These relationships have been very well studied in the robotics community as a key aspect of system characterization (Sciavicco and Siciliano, 2005). However, differential kinematics are largely unstudied with respect to speech motor control. It was suggested by Saltzman et al. (2006) that differential kinematics could be used to quantify the debate over the nature of speech production goals. However, we are not aware of any studies which specifically attempt to model the differential kinematics of speech production, nor do we know of any that utilize the differential kinematics for the purposes of characterizing the speech production system. We demonstrate the accuracy of Artificial Neural Networks (ANNs) and Locally-Weighted Linear Regression (LWR) by evaluating them on data relevant to speech production. We also argue for the utility of looking at speech data in this way.

Differential kinematics offer an exciting new way of looking at speech data, with the potential to offer many insights and to be useful for many applications. They form a rigorous mathematical framework for exploring systematic characterization of the speech production

system in several respects. For instance, they allow for comparison of the degrees of freedom in the articulator space versus the task space (i.e., redundancy). They can facilitate the identification of localized reductions in task-space degrees of freedom (i.e., singular postures) which may occur, for instance, when articulators become perfectly aligned. They also provide a basis for inverse kinematic algorithms, deriving the equations of motion for a system, looking at the force control and interface interaction (e.g., interaction between the tongue and palate), and designing and evaluating task-space control schemes.

Differential kinematics may be of particular use in studying speech due to the longstanding debate over the nature of speech tasks – i.e., whether they are acoustical (e.g., Guenther (1994)) or articulatory (e.g., Saltzman and Munhall (1989)). This debate has been encapsulated in specific models of motor coordination, including the DIVA model (Guenther, 1995; Guenther et al., 1998) and the Task Dynamics account (Saltzman and Munhall, 1989; Saltzman and Byrd, 2000), which are based on different assertions about how best to describe the task space. Using differential kinematics, one can apply computational methods such as the Uncontrolled Manifold Method (UCM) (Scholz and Schöner, 1999), recently suggested by Saltzman et al. (2006) as a way to quantify the debate concerning the nature of the speech production tasks. The UCM takes advantage of the fact that, using the differential kinematics, one can divide the observed kinematic variability into that which is relevant to a given task, and that which is not. Given two competing task descriptions, the better one will show a higher proportion of variability in the task-relevant portion.

The potential also exists for differential kinematics to inform new methods for acoustic-to-articulatory inversion – i.e., the problem of recovering vocal tract configurations given only speech acoustics. Early, analytical approaches to this problem showed the apparent redundancy in the system, which implies that the forward map is non-invertible (Mermelstein and Schroeder, 1965; Wakita, 1973). Later techniques attempted to resolve this ambiguity, either by using statistical properties of the map (Atal and Rioul, 1989; Papcun et al., 1992; Hogden et al., 1996; Qin and Cerreira-Perpiñán, 2007; Lammert et al., 2008; Ananthakrishnan et al., 2009; Qin and Cerreira-Perpiñán, 2010; Ghosh and Narayanan, 2010) or through analysis-by-synthesis (Atal et al., 1978; Boë and Bailly, 1992; Schroeter and Sondhi, 1994; Panchapagesan and Alwan, 2011). Statistical methods similar to those explored here have been utilized to learn the inverse map directly (Rahim et al., 1991; Richmond, 2010; Al Moubayed and Ananthakrishnan, 2010).

While these inversion efforts have yielded substantial progress, differential kinematics can offer new insight by way of identifying and quantifying redundancy in the system. More importantly, new inversion techniques should be possible which confront the non-uniqueness problem in novel ways. Many iterative, computational methods have been developed in the robotics community for finding a reasonable path through articulator space to reach a position in task space, even despite nonuniqueness. Some common solutions utilize the differential kinematics to accomplish this, including Jacobian pseudoinverse methods (Whitney, 1969), the Jacobian transpose method (Balestrino et al., 1984; Wolovich and Elliot, 1984) and damped least squares methods (Nakamura and Hanafusa, 1986; Wampler, 1986).

In order to estimate the direct and differential kinematics of the speech production system, one must choose a statistical method that is capable of (1) estimating complex, nonlinear relationships to a high degree of accuracy, and (2) facilitating the extraction of partial derivatives that make up the differential kinematics, preferably directly and without appealing to numerical approximations. These are the key technical challenges for any candidate method. In addition, it is desirable to have a method which is easily designed and trained. These practical challenges are of equal importance for the utility of a chosen method. We explore the use of ANNs and LWR, two methods which represent drastically different underlying assumptions in terms of model fitting and prediction.

ANNs were recently suggested for this purpose by Saltzman et al. (2006). Their abilities as universal function approximators allows them to learn complex maps (Cybenko, 1989; Hornik et al., 1989). However, ANNs have some practical drawbacks, in the sense that they are notoriously difficult to design and slow to train (Bishop, 2006; Wilamowski et al., 2008). Training is computationally expensive, involving optimization of a non-convex objective function, and it can be difficult to determine training convergence in practice. If new data arrive, training must be repeated.

LWR offers a complementary set of advantages to ANNs. It has many practical advantages, most notably efficiency of training and the presence of fewer free parameters. The method is powerful enough to approximate very complex functions, even despite assumptions of local linearity. Indeed, local linearizations are ubiquitous in kinematics and control applications because they are entirely appropriate, practical approximations

to the kinds of nonlinearities seen in many motor systems. The relationship between velocities in articulator and task space is often expressed mathematically as a linear transformation (see Equation 3, below).

We have previously reported on an initial effort to estimate kinematic relationships from data (Lammert et al., 2010). This paper constitutes an expansion of that work, providing a substantial refinement of the techniques previously presented. To that end, we have refined our previous formulation of LWR. We have also implemented a heldout data validation scheme for parameter optimization with respect to both algorithms. This has resulted in more confidence about the generalizability of the results and, crucially, in superior estimation accuracies according to our previous evaluation metrics. We have also included additional evaluations to assess our modeling of kinematic aspects which were neglected in previous work. Additionally, we have tested our methods on an additional data set, which was designed to more accurately reflect data which might be acquired in real speech studies. Finally, we present a more thorough discussion of the motivations for kinematic analysis, as well as its potential applications for studying speech production.

Section 2 will serve to explain our methodology, including a formal description of direct and differential kinematics in Section 2.1, the creation of our data sets in Sections 2.2 and 2.3, as well as a review of ANNs and LWR in Sections 2.4 and 2.5. An explanation of our model optimization procedure is in Section 2.6 and an overview of our experiments for evaluation is in Section 2.7. In Section 3, we provide the results of our experiments: Section 3.1 provides the results on two synthetic data sets in terms of accuracies in estimating direct and differential kinematic relationships, while Section 3.2 provides direct kinematic accuracies on a real speech data set. In Section 4, we discuss the performance of ANNs and LWR, and in Section 5, we present our concluding remarks and remaining challenges for this line of work.

2. Methods

Broadly stated, our methodology involves the generation of two large corpora of parallel articulator and task vectors using an articulatory model of the vocal tract. We then employ both ANNs and LWR to estimate the direct and differential kinematics of this model from the data. Parameter tuning was done by examining accuracy of the direct kinematic estimation on a development set of heldout data. Evaluation was then performed by examining accuracy of both the direct and differential kinematics on a heldout test set. In this section, we provide detail about the theory and implementation of our methods, as well as the specific rationale for our choices.

We would like to be especially clear regarding the motivation for using synthetic data, given that this study is ultimately aimed at developing methods which can be used on real data. Indeed, we intend to apply these methods to real data, which we discuss in our concluding remarks (see Section 5). The use of synthetic data is driven by the need to evaluate the differential kinematic estimates. These cannot be evaluated on real data since the relevant relationships are not directly observable, in contrast to direct kinematics which can be precisely measured and for which the accuracy can be evaluated by calculating the residual error in task space. For synthetic data, on the other hand, the differential kinematics are known *a priori* and can be used as a standard for evaluation.

With this in mind, we also develop and demonstrate methods for parameter tuning (i.e., model selection) and model training that do not depend on knowing the differential kinematics, and therefore can be performed on real speech data where knowledge of these aspects is not available. We also wish to emphasize that deriving estimates of the direct and differential kinematics in the way we demonstrate is completely repeateable on real articulatory data.

2.1. Direct and Differential Kinematics

Given a vector q, representing n low-level articulator variables of the system, and a vector x, representing m high-level task variables of the system, the relationship between them is commonly expressed by the direct kinematics equation, of the form:

$$x = k(q) \tag{1}$$

where the function $k(\cdot)$ represents the forward map which is assumed to be complex and nonlinear. It is the forward map that we are attempting to learn from a representative corpus of parallel example pairs of q and x.

We are particularly interested in modeling $k(\cdot)$ so as to facilitate derivation of the Jacobian matrix:

$$J(q) = \begin{pmatrix} \partial x_1/\partial q_1 & \cdots & \partial x_1/\partial q_n \\ \vdots & \ddots & \vdots \\ \partial x_m/\partial q_1 & \cdots & \partial x_m/\partial q_n \end{pmatrix} \tag{2}$$

The Jacobian is a compact representation of knowledge regarding the posture-specific 1st-order partial derivatives of tasks with respect to articulators. It allows us to

write the differential kinematics equation, which relates articulator velocities and task velocities in the following way:

$$\dot{x} = J(q)\dot{q} \tag{3}$$

Note that this equation expresses the relationship between $\dot{q}$ and $\dot{x}$ as a linear transformation. This kind of approximation allows for such an elegant mathematical formulation of the kinematics. It also highlights why locally-linear methods, such as LWR, are appropriate for estimating kinematic relationships.

It is possible that further approximations, if appropriate, may allow additional simplification of the mathematics. In certain applications (e.g., Cootes et al. (2001)), it is appropriate to assume that the differential kinematics are globally constant. This means that the Jacobian is no longer a function of the pose (i.e., $J(q) \approx J$ and $\dot{x} = J\dot{q}$). We do not believe that this approximation is appropriate for the speech production system, but future work may help determine this (see Section 5).

As mentioned above (see Section 1), an understanding of differential kinematics allows one to address fundamental issues of system characterization. Indeed, the Jacobian is one of the most useful tools for characterizing systems. For example, postural singularities can easily be identified by examining the rank of the Jacobian. If it is rank-deficient, then there is a singularity at the current posture. Similarly, an examination of the range and null of the Jacobian provides a formal analysis of articulatory redundancy. The Jacobian also represents sufficient information about the kinematics to apply the Uncontrolled Manifold Method.

2.2. Kinematic Model

The model we utilize in generating our data is the TAsk Dynamic Application (TADA) (Nam et al., 2004, 2006), an implementation of the Task Dynamic model of articulatory control and coordination (Saltzman and Munhall, 1989). The specific articulatory model at the core of TADA is the Configurable Articulatory Synthesizer (CASY) (Rubin et al., 1996; Iskarous et al., 2003), which specifies geometric articulator variables.

As outlined at the beginning of this section, we are using synthetic articulatory data to facilitate evaluation of the differential kinematics. We will not make any bold claims about the realism of TADA as a model for speech production, although we believe that is remains faithful to real speech articulation in certain key ways. Evidence for this assertion comes from work using TADA for analysis-by-synthesis (Mitra et al., 2009; Nam et al., 2010; Mitra et al., 2010, 2011). Articulatory time series extracted from this process have been shown to improve automatic speech recognition, which would be unlikely if correspondence was poor between TADA's articulation and real speech.

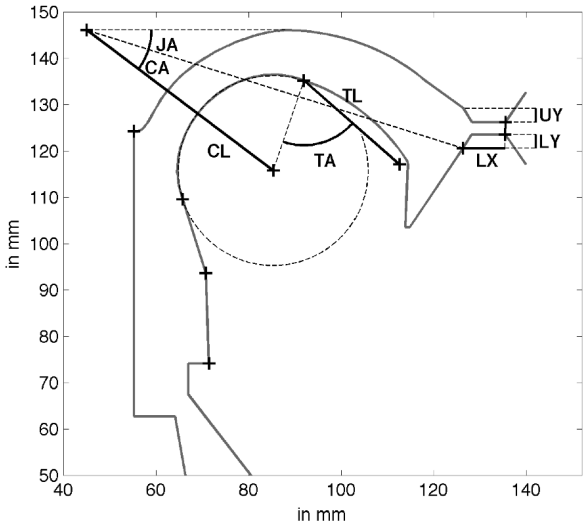

Figure 1: *A visualization of the Configurable Articulatory Synthesizer (CASY) in a neutral position, showing the outline of the vocal tract model (gray line). Overlain are the key points (black crosses) and geometric reference lines (dashed lines) used to define the articulatory parameters (black lines and angles), which are also labeled.*

We note that CASY/TADA is only one possible implementation of a kinematic model for speech production. The crucial aspect for our purposes is only that we use a model that is well-informed about the kinds of nonlinearities that are relevant for speech. The forward model incorporated in TADA is such a model. It is based on the geometrical equations of Mermelstein (1973) for describing vocal tract congurations and deriving vocal tract outlines from articulator positions. Assuming a simple parameterized shape for the palate and rear pharyngeal wall, the constriction task variables can also be derived from the outline.

The forward model implemented in CASY uses geometric articulator variables which are shown in Figure 1. These include lip protrusion (LX) and vertical displacements of the upper lip (UY) and lower lip (LY) relative to the teeth. The jaw angle (JA) and tongue body angle (CA) are defined relative to a fixed point above and behind the velum, as is the tongue body length (CL). The tongue tip is defined by its length (TL) from the base to the tip, as well as its angle (TA) relative to the tongue body center and the posterior base of the tongue blade.

Using the vocal tract outline, it is possible to calcu-

late a vector of task variables for each articulator configuration. These tasks are articulatory in nature, but are high-level relative to the articulator variables. The tasks are as follows: lip aperature (LA) and protrusion (PRO), tongue body constriction degree (TBCD) and location (TBCL), as well as tongue tip constriction degree (TTCD) and location (TTCL).

The complexity of the model varies depending on the task variable in question. All tongue-related tasks are quite complex and nonlinear, with the most prominent nonlinearities conforming to trigonometric and polynomial functions. The tongue tip tasks are the most complex, due to many nonlinearities and to having the greatest articulatory redundancy (five different articulators contribute to them). At the other extreme, the lip protrusion task variable depends only on the lip protrusion articulatory variable in a linear way (i.e., a linear, non-redundant relationship). All other task variables are related to the articulators by at least two nonlinearities. The full list of equations is given in Appendix A.

TADA is an implementation of precisely those dynamical equations described by Saltzman and Munhall (1989) in their dynamical approach to control of speech production gestures. A full exposition of the model can be found in referring to Appendix 2 of that publication. In brief, it states that the equation of motion for the actively controlled articulators is:

$$\ddot{q}_A = J * (M^{-1}[-BJ\dot{q} - K\Delta x(q)]) - (J * \dot{J}\dot{q} + (I_n - J * J))\ddot{q}_d \quad (4)$$

where $\ddot{q}_A$ represents the articulatory acceleration vector, which encapsulates the active driving influences on the model articulators. Also, $\ddot{q}_d = B_n\dot{q}$ is the acceleration damping, where B_n is a diagonal matrix of damping constants. The other variables include M, a diagonal matrix of inertial coefficients, B, the diagonal matrix of damping coefficients for the task variables, K, the stiffness coefficients for the task variables, and J, the Jacobian matrix.

2.3. Data Sets

We generated two large codebooks of data using the CASY/TADA forward model. These two codebooks provide complementary ways of populating the articulator space. We first uniformly populated the articulator space (i.e., data points conformed to a grid within the space). The filled space was bounded by the articulator ranges obtained from the speech-relevant data (see below and Table 1). Within those bounds, we defined a rectangular grid with 5 evenly spaced points along each of the 8 dimensions. This created 5^8 = 390,625 total articulator vectors and their accompanying 390,625 task vectors of length 6. This uniformly-distributed codebook represents all possible vocal tract configurations with equal density of coverage. Building this data set provides a way of testing our estimation methods that is independent of TADA's ability to move CASY's articulators in a realistic way. It also provides broad coverage of the articulator space, which should lead to thorough testing of the modeling methods.

Articulatory Variable	min	max	range	units
Lip Protrusion (LX)	9.11	12.00	2.89	*mm*
Jaw Angle (JA)	1.11	1.41	0.31	*rad*
Upper Lip Displacement (UY)	−4.78	0.93	5.70	*mm*
Lower Lip Displacement (LY)	−8.57	19.95	28.52	*mm*
Tongue Body Length (CL)	68.59	83.62	15.03	*mm*
Tongue Body Angle (CA)	−0.36	0.04	0.40	*rad*
Tongue Tip Length (TL)	6.50	44.38	37.88	*mm*
Tongue Tip Angle (TA)	−0.24	1.40	1.67	*rad*

Table 1: *Ranges of the various CASY articulator variables, as observed during synthesis of the speech-relevant data set described in Section 2.3.*

It seems very unlikely, however, that real articulatory recordings of natural speech would reflect this kind of data distribution. We would expect that real speech would fill only a subspace of the full articulator space due to inter-articulator correlations and the likelihood that certain configurations are not used for speech. Thus, it is equally important to evaluate estimation accuracy on a codebook that is representative of real speech data which might be acquired from speaking subjects.

For this reason, we used CASY/TADA to synthesize a set of English sentences. We used TADA to synthesize 30 English sentences taken from the MOCHA-TIMIT corpus (Wrench and Hardcastle, 2000) at an effective sampling rate of 200 Hz. The specific sentences constitute the first 30 sentences of that corpus, and have been reproduced in Appendix B. For reference, the resulting articulator ranges are shown in Table 1. This synthesis provided us with speech-relevant codebook containing 17,198 total articulator vectors of length 8. These were accompanied by the same number of task vectors of length 6.

From these two large codebooks, we created data sets for evaluation by randomly sampling vectors without replacement. Data sets were of sizes 78, 156, 312, 625, 1250, 2500 and 5000. Having data sets of standard sizes facilitated a direct comparison between both algorithms. The reason for producing data sets of varying sizes was to determine (1) whether the amount of training data affected the accuracy of estimation (i.e., training effects)

and (2) whether the overall amount of training data was sufficient (i.e., convergence of performance). Although the largest data set seems modest in size, we were able to show that it is sufficiently large (see Section 4).

2.4. Artificial Neural Networks

In keeping with the suggestion of Saltzman et al. (2006), we implemented a directed, multilayer, feedforward neural network, otherwise known as a multilayer perceptron (MLP) (Bishop, 2006). We acknowledge that many alternative ANN architectures and topologies exist, from mixture of experts networks (Jacobs et al., 1991; Jordan and Jacobs, 1995) to fully connected networks (Wilamowski et al., 2008). These are equally capable of learning arbitrarily complex functions. However, the advantage of using the MLP architecture is in the ability to easily and analytically (i.e., without the need for numerical methods) extract the Jacobian. The method for doing this is described below.

Our network hidden nodes employed the commonly-used sigmoidal transfer functions. It is possible to tailor ANNs to a specific estimation task by choosing different nonlinear transfer functions, especially if the functional form of the map is known *a priori*. Many different functions can be used while still maintaining the universal approximation power of MLPs (Duch and Jankowski, 1999). Moreover, as long as these nonlinearities are differentiable, it is still possible to easily extract the Jacobian. Our intention here was to choose a transfer function that offers the most generality, so as to make minimal assumptions about the structure of real speech data. When it comes to applying these techniques on real speech data, this kind of generality and flexibility will hopefully be a benefit.

We trained the ANNs with the standard error backpropagation (Rumelhart et al., 1986b,a; Jordan, 1992). We acknowledge that more efficient training algorithms have been suggested for ANNs with an MLP-type topology. The Levenberg-Marquardt algorithm (Hagan and Menhaj, 1994; Toledo et al., 2005; J.-M., 2008) is widely considered to be the most efficient algorithm for training MLP architectures. Still, there is no guarantee that these algorithms provide greater accuracy after training. Since our primary concern is accuracy and not speed of training, we did not implement these methods.

The error function that backpropagation attempts to minimize through gradient descent is a standard least squares function of the form:

$$E_n = \frac{1}{2} \sum_k (\hat{x}_{nk} - x_{nk})^2 \tag{5}$$

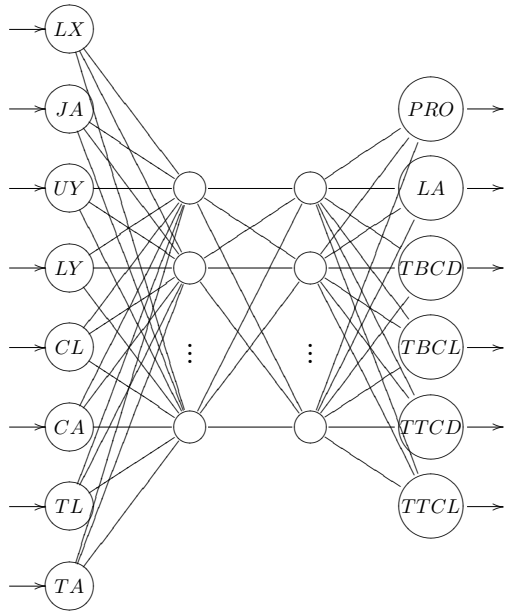

Figure 2: *Our ANN topology, an instance of a fully-connected, feedforward Multilayer Perceptron with two hidden layers. The number of nodes in the hidden layers is a free parameter which determines the complexity of the model, after training.*

where the quantity $\hat{x}_{nk}$ is the output of the network at unit k when presented with input data point n.

Our network topology is represented in Figure 2. This network had linear input and output nodes, corresponding to each of the articulatory and task variables, respectively. Between were two layers of hidden units, all with sigmoidal activation functions:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

Upon completion of training, a pose-specific Jacobian matrix can be obtained for any articulatory input vector. This can be done with the use of numerical methods (Bishop, 2006). However, the Jacobian can also be obtained analytically for a network of this kind using the feedforward formalism (Jordan and Rumelhart, 1992; Saltzman et al., 2006). Note that each hidden node has a sigmoidal activation function, we can write the derivative of each node's activation with respect to its input as follows:

$$\delta = z(1 - z) \tag{7}$$

Then, we can arrange these values in a diagonal matrix, denoted Δ_i. The Jacobian for a given input posture

is then

$$J = \Delta_{out} W_{out,H2} \Delta_{H2} W_{H2,H1} \Delta_{H1} W_{H1,in} \quad (8)$$

where W_{ij} is the weight matrix connecting layer i to layer j.

General design parameters of networks with this architecture are many, and include the learning rate (α), the number of training iterations (num_{iter}), the number of hidden layers and the number of nodes in each hidden layer. It is possible in theory to tune each of these parameters to optimal values, but the tuning procedure becomes intractable with so many free parameters. The complexity of the model is essentially determined by the number of nodes in each hidden layer, which we denote p_{ANN}. Thus, we chose to treat only this as a free parameter within the scope of parameter tuning for ANNs. Other design parameters were fixed, so that $num_{iter} = 300$, $\alpha = 0.001$, and the number of hidden layers was always 2. Limiting the free parameters in this way still allows for proper tuning to be performed, so as to promote generalizability of the statistical model (see Section 2.6), while keeping this procedure tractable.

2.5. Locally-Weighted Regression

Locally-weighted linear regression is one outcome of a long line of research into nonparametric methods which use locally-defined, low-order polynomials to approximate globally nonlinear functional relationships. Much of this early work is contained in the Statistics literature, where these techniques have a long successful history (Cleveland, 1979; Cleveland and Devlin, 1988; Cleveland et al., 1988). Atkeson et al. (1997) surveyed much of the early work on this topic from the Statistics literature, and also provided a unifying view of these techniques for the Machine Learning community.

LWR is a memory-based, lazy learning method, which means that it keeps the entire data set in memory and uses it directly at prediction time in order to calculate the parameters of interest. Formulation of this technique begins by assuming that the data were generated by a model following

$$x_i = k(q_i) + \epsilon \quad (9)$$

where k is a function which can be nonlinear, in general. The value ϵ represents the noise which is assumed to follow a Gaussian distribution

$$\epsilon \sim N(0, \sigma^2) \quad (10)$$

a normal distribution with mean 0 and variance σ^2.

We would like to fit the data in a local region defined by the data point q_i. The measure of locality K is taken to be a Gaussian kernel function

$$K(q_i, q_j, h) = exp\{-(q_i - q_j)^T H (q_i - q_j)\} \quad (11)$$

although any such kernel can be utilized. H is a positive semi-definite diagonal matrix, with diagonal elements equal to $1/2h^2$. The value h is a free parameter with a straightforward interpretation: it is the standard deviation of the Gaussian kernel. If h has the same value in each column in H (i.e., in all directions in articulatory space), the kernel will be spherical. Since the articulators in this case have a variety of ranges and units, we chose to set h differently in each direction, which gives one value for each articulator variable.

We assume that a linear model is an appropriate approximation to the forward map within the local region. Thus, the model we would like to fit locally is of the form

$$x_i = \beta_i^T q_i \quad (12)$$

where β is the vector of regression coefficients.

The error function that needs to be minimized is an extension of the standard weighted least squares function of the form:

$$E_i = \frac{1}{2} \sum_j [(\beta_i^T q_i - x_i)^2 K(q_i, q_j, h)] + \frac{\lambda}{2} ||\beta|| \quad (13)$$

The second term is a regularization term, which contains the ridge regression parameter γ. In cases when there are very few data points near q_i, a danger is that the regression matrix may become nearly singular and numerical issues will arise in computing the solution. Adding the regularization term prevents this problem at the expense of biasing the solution very slightly. In practice, the parameter γ can be effective even when $\ll 1$, which ensures a marginal bias of the solution.

An analytical solution can be found for β with ridge regression as follows:

$$\beta_i = (Q^T W_i Q + \gamma I)^{-1} Q^T W_i X \quad (14)$$

The matrix, W_i, is a diagonal weight matrix, formed from the outputs of the kernel function with a fixed q_i.

An illustrative example is shown in Figure 3 for a toy set of low-dimensional nonlinear data. The Gaussian kernel is visualized along with a locally-linear model for one point in articulator space A global fit line, also visualized, can be obtained as a agglomeration of many locally-linear models. This is done by simply identifying an arbitrary number of points in articulator space

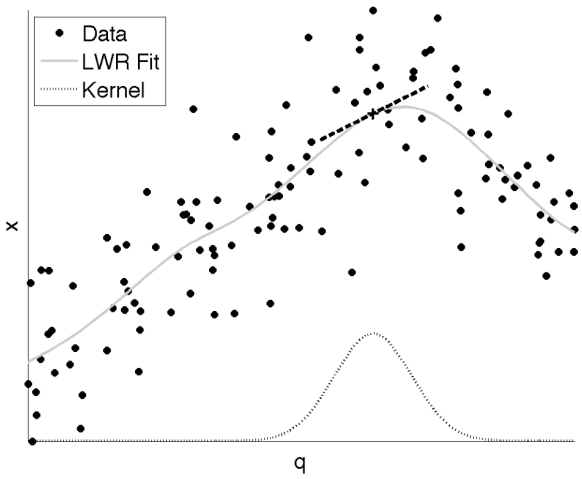

Figure 3: *An illustration of modeling with LWR. For a particular point (black cross) a local region is defined in articulator space by a Gaussian-shaped kernel (gray dashed curve). A line is fit in the local region using a weighted least-squares solution, indicated by the black dashed line. The global fit is generated by repeating this procedure at a large number of local regions. The resulting fit can be quite complex (gray curve), and depends on the width of the kernel.*

and solving for β_i and x_i at each of them. The locally-linear solutions can be said to constitute a global fit to the data.

Obtaining the Jacobian from this model is trivial, since it is already linear. The regression vector β contains the locally-relevant partial derivatives. In other words, the values of this vector are the elements of the Jacobian.

The only design parameters for LWR are the kernel width parameters, h. The values determine the complexity of the resulting model, much like the number of hidden nodes in the ANN. Therefore, these parameters must be subject to careful tuning in the evaluation experiments. While in principle, h can vary freely in each direction of articulator space, this would make the tuning search space quite large. At the same time, we can assume that the kernel width will be similar in each direction if only it is normalized by the articulator range. Thus, we define a hyperparameter p_{LWR}, such that, for articulator j, the kernel width is $h_j = p_{LWR} \cdot range(Q_{*,j})$. This heuristic approach allows us to collapse the parameter tuning process into a 1-dimensional problem of tuning the hyperparameter. Note that a more complex model is indicated by *smaller* values of p_{LWR}, the opposite of p_{ANN}.

2.6. Model Selection

An important practical challenge in applying these techniques is determining good values for the free parameters p_{ANN} and p_{LWR}. This is the problem of model selection, which is common to many statistical analyses or model fitting applications. For instance, if we are planning to use LWR on a given data set in order to estimate the Jacobian at a posture of interest, it is necessary to select a value for p_{LWR} which provides a close fit, but which promotes generalizability by avoiding overfitting. Finding this critical value is commonly known as the bias-variance tradeoff.

Optimization of the parameter value can be done using a heldout data validation procedure, whereby the data set is partitioned into a training and development set. Using a wide range of values for the free parameter, the training set is used to provide an estimate for each data point in the development set. The parameter value which provides the highest accuracy is selected as the optimal value. Heldout data validation is a principled method for model selection. However, we note that a rotating heldout validation procedure (e.g., n-fold or leave-one-out cross-validation), while deemed excessive in this situation, would be more robust on real data.

We implemented this kind of model selection procedure by randomly assigning 90% of the vectors in a given data set to a training subset and 5% to a development subset. The remaining 5% were set aside in a separate test set, to function as our postures of interest for the purposes of evaluation later on. Performance on the development set, given a particular parameter value, was determined by calculating the mean across all tasks of the normalized root mean squared error (RMSE). Note that the differential kinematics are not used at all in the model selection process. Model selection can be done in terms of the direct kinematics. There is, therefore, no reason why this procedure cannot be repeated on entirely real speech data, when the Jacobian is not available for the purposes of development.

For the uniformly-distributed data set, it was determined that the optimal network topology had hidden layers containing 200 nodes each (i.e., $p_{ANN} = 200$). The optimal value for p_{LWR} was determined to be 0.300, corresponding to a Gaussian kernel with standard deviation equal to 30.0% of each articulator's range. For the speech-relevant data set, the optimal network had $p_{ANN} = 250$, and LWR was optimal with $p_{LWR} = 0.067$, corresponding to a Gaussian kernel with standard deviation equal to 6.7% of each articulator's range. It is notable that there was a shift in parameter values between the two data sets, such that models were allowed to become more complex on the speech-relevant

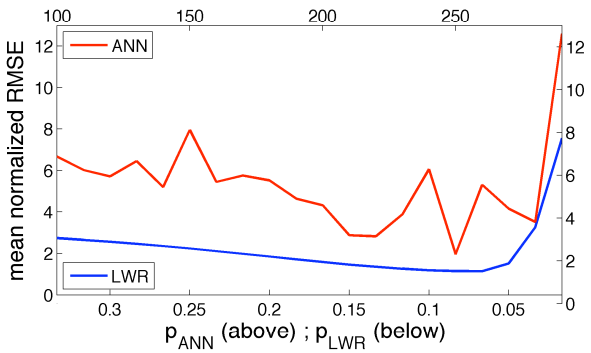

Figure 4: *A comparison of LWR and ANN performance on held-out speech-relevant data as a function of their free parameter values. Each displays a minimum point, which represents the optimal value for that parameter. Choosing other values runs the risk of either overfitting or underlearning. Note that the axis values for p_{LWR} are reversed, so that the right-hand side indicates more complex models for both parameters.*

data. This is likely due to the speech-relevant data being more densely packed in the articulatory space. This allows more detail about the kinematics to be obtained. The relatively sparser uniformly-distributed data requires more generalization in order to avoid overfitting.

To illustrate the model selection procedure, the performance of both algorithms are shown in Figure 4 over a range of parameter values on a speech-relevant development set. It should be noted that the axis values for p_{ANN} and p_{LWR} are reversed with respect to each other. This was done to consistently indicate more complex models (represented by large values of p_{ANN} and small values of p_{LWR}) toward the right-hand side of the plot. For LWR, the errors form a smooth and apparently convex function with a unique minimum over the range of explored values. This inspires confidence that the parameter value selected is the optimal one. The errors for the ANNs suggest a similar trend, but with a much more jagged appearance. This most likely reflects the difficulties associated with training ANNs. That is to say, backpropagation likely got stuck in a local minimum and did not successfully converge on the global minimum of the objective function. Still, the overall shape of the tuning curve is highly consistent with that displayed by LWR.

2.7. Evaluation

After performing model selection, we used the optimal parameter values to perform a final evaluation on the test set (see Section 2.6 for a description of the development and test sets). The entire procedure, including model selection and evaluation, was done identically

Direct Kinematic Accuracy – Uniform Data

Task Variable	RMSE_{ANN}	RMSE_{LWR}
PRO	2.78 mm (92.6 %)	0.06 mm (2.1 %)
LA	3.13 mm (4.8 %)	0.13 mm (0.2 %)
TBCL	24.14 deg (12.8 %)	20.37 deg (10.8 %)
TBCD	4.38 mm (13.3 %)	2.01 mm (6.1 %)
TTCL	12.85 deg (4.1 %)	20.69 deg (6.6 %)
TTCD	4.34 mm (4.5 %)	2.03 mm (2.1 %)

Direct Kinematic Accuracy – Speech-Relevant Data

Task Variable	RMSE_{ANN}	RMSE_{LWR}
PRO	0.40 mm (13.8%)	0.05 mm (1.6%)
LA	0.38 mm (1.2%)	0.31 mm (1.0%)
TBCL	2.00 deg (2.3%)	1.08 deg (1.2%)
TBCD	0.36 mm (2.0%)	0.18 mm (1.0%)
TTCL	0.67 deg (0.9%)	0.42 deg (0.5%)
TTCD	0.67 mm (2.0%)	0.40 mm (1.2%)

Table 2: *Accuracies of the direct kinematic estimates for both algorithms on each data set. Displayed are the root mean squared error (and RMSE as a percentage of the task range) across all vectors in the test sets.*

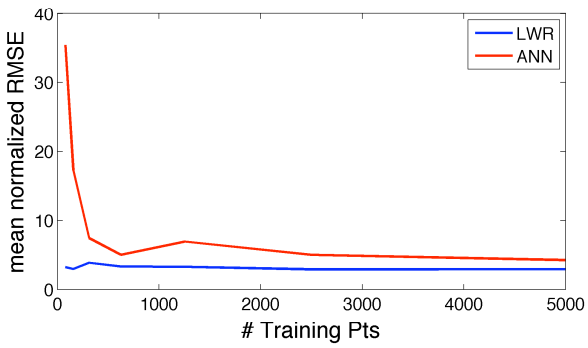

Figure 5: *A comparison of LWR and ANN performance across a variety of data set sizes. It is clear that LWR outperforms ANN no matter the quantity of training data. Moreover, LWR is very insensitive to the amount of training data, unlike the ANN. Moreover, both algorithms seem to plateau around 3000 data points, indicating that the amount of training data we used is sufficient.*

on uniformly-distribute and speech-relevant data sets of all sizes.

3. Results

3.1. Synthetic Data

Our experimental results were evaluated with respect to the accuracy of modeling the direct and differential kinematics. Methods for quantifying this accuracy can be found by inspecting the direct and differential kinematics equations, as written in Equations 1 and 3, though assessment of each is done differently.

Since an accurate model of direct kinematics should constitute the forward map, $k(q)$, accuracy can be defined in terms of its prediction of the task vector, x,

Diff Kin Accuracy – ANN, Uniform Data

J_i	r	$\angle(J_{A,i}, J_{N,i})$	$\parallel J_{A,i} \parallel$	$\parallel J_{N,i} \parallel$
J_{PRO}	0.35	64	1.00	1.22
J_{LA}	0.89	27	1.77	2.08
J_{TBCL}	0.65	39	9.21	5.66
J_{TBCD}	0.74	42	1.04	1.11
J_{TTCL}	0.74	40	4.66	3.80
J_{TTCD}	0.71	37	1.77	1.87

Diff Kin Accuracy – LWR, Uniform Data

J_i	r	$\angle(J_{A,i}, J_{R,i})$	$\parallel J_{A,i} \parallel$	$\parallel J_{R,i} \parallel$
J_{PRO}	1.00	0	1.00	0.88
J_{LA}	1.00	1	1.77	1.75
J_{TBCL}	0.70	35	9.21	6.19
J_{TBCD}	0.90	23	1.04	0.84
J_{TTCL}	0.76	35	4.66	3.63
J_{TTCD}	0.95	14	1.77	1.56

Table 3: *The mean correlation coefficient (r) between the rows of the analytically-derived Jacobian (J_A) and the Jacobian estimated with the ANN (J_N) and LWR (J_R) from the uniformly-distributed data set. Also shown are the mean angle in degrees, the norm of each vector. The abbreviations that instantiate the subscript i of J correspond to the task variables described in Section 2.2*

Diff Kin Accuracy – ANN, Speech Data

J_i	r	$\angle(J_{A,i}, J_{N,i})$	$\parallel J_{A,i} \parallel$	$\parallel J_{N,i} \parallel$
J_{PRO}	0.76	43	1.00	0.29
J_{LA}	0.79	38	1.78	1.45
J_{TBCL}	0.95	15	9.33	4.31
J_{TBCD}	0.94	20	1.03	1.04
J_{TTCL}	0.95	15	4.86	2.50
J_{TTCD}	0.89	22	2.09	1.70

Diff Kin Accuracy – LWR, Speech Data

J_i	r	$\angle(J_{A,i}, J_{R,i})$	$\parallel J_{A,i} \parallel$	$\parallel J_{R,i} \parallel$
J_{PRO}	0.84	27	1.00	0.09
J_{LA}	0.84	32	1.78	1.14
J_{TBCL}	0.89	25	9.33	3.81
J_{TBCD}	0.94	18	1.03	0.76
J_{TTCL}	0.96	15	4.86	2.12
J_{TTCD}	0.88	23	2.03	1.27

Table 4: *The mean correlation coefficient (r) between the rows of the analytically-derived Jacobian (J_A) and the Jacobian estimated with the ANN (J_N) and LWR (J_R) from the speech-relevant data set. Also shown are the mean angle in degrees, the norm of each vector. The abbreviations that instantiate the subscript i of J correspond to the task variables described in Section 2.2*

given an articulator vector, q. This can be quantified in terms of the root mean square error (RMSE) for predicting all task vectors in the test set. The task-wise accuracies are shown in Table 2 for both algorithms on the two 5000-point data sets. Also displayed are the normalized RMSE values, reflecting the errors as a percentage of the observed task range. Figure 5 shows a comparison of LWR and ANN performance on data sets of varying sizes drawn from the uniformly-distributed codebook.

Accuracy in estimating the differential kinematics can be assessed directly in terms of the Jacobian matrix, $J(q)$, for any articulatory configuration, q. In turn, this can be judged by comparing the relative contributions of all articulators to each task. Mathematically, this is done by calculating the vector angle between each row of the estimated Jacobian with its corresponding row in the analytically-derived Jacobian from CASY/TADA. Relatedly, and perhaps more interpretably, one can quantify this using Pearson's r. Tables 3 and 4 display the average angle and correlation values for both algorithms operating on both 5000-point data sets. Also displayed are the row-wise Euclidean vector norms of the estimated Jacobian rows and the analytically-derived Jacobian. Comparing these values gives an indication of how well the overall magnitude of the articulator contributions was estimated for each task.

Direct Kinematic Accuracy – Real Data

Task Variable	$\mathbf{RMSE}_{ANN}$	$\mathbf{RMSE}_{LWR}$
LA	2.32 mm (13.7 %)	1.66 mm (9.8 %)
$TTCD$	3.04 mm (9.6 %)	2.96 mm (9.4 %)
$TBCD$	2.61 mm (14.8 %)	2.33 mm (13.2 %)

Table 5: *Accuracies of the direct kinematic estimates for both algorithms on a vocal tract data set acquired using rtMRI from an American English speaker reading 65 TIMIT sentences. Displayed are the root mean squared error (and RMSE as a percentage of the task range) across all vectors in the test sets.*

3.2. Real Speech Data

In preliminary experiments, we have applied the proposed methods to real speech data (Lammert et al., 2011). In particular, we applied these methods to a subset of the real-time MRI (rtMRI) data presented in Narayanan et al. (2011) from one speaker of American English (5000 data points = 65 sentences at 23.33 frame/second). We extracted articulator and task variables, corresponding to those described above (see Section 2.2), from midsagittal vocal tract outlines automatically fit to the rtMRI images (Bresch and Narayanan, 2009). Outlines fit with this method are pre-segmented into anatomical structures (e.g., tongue, lips, etc.) that facilitate the extraction of articulator and task variables. Articulator variables were extracted using geometrical relationships between these outlines. For instance, jaw angle was defined as the angle between the jaw outline

and pharyngeal wall outlines. Constriction degree task variables were extracted by calculating the Euclidean distance between the closest points on the relevant outlines. For example, lip aperture was defined as the minimum distance between the upper and lower lip outlines.

Direct kinematic accuracies were quantitatively evaluated as before, and the results are presented in Table 5. Results show that constriction degrees were accurate to 2.7 mm on average. This level of accuracy is close to the 2.9 mm pixel width (i.e., spatial resolution) of the rtMRI data set. Moreover, results are consistent with the results on synthetic data, in that LWR outperforms ANN on this real speech data, as well.

4. Discussion

The overall best estimation was observed by LWR operating on the speech-relevant data set. Errors in estimating the direct kinematic relationships in this situation are approximately 1% of the task range, with only slight variability across tasks. The estimates of the differential kinematics in this situation are quite accurate, with all Jacobian rows displaying correlation coefficients of approximately 0.90 with the ground truth, again with slight variation across tasks. We believe that these accuracies are sufficiently good as to be useful in facilitating a variety of kinematic analyses for characterizing speech production.

We observe that LWR outperforms the ANN in terms of accuracy, for the map explored here and for these particular parameters values. We must note that this is not the first time that locally-linear models have been shown to outperform ANNs in practice (Lawrence et al., 1996). The difference in accuracy most likely reflects practical difficulties associated with selecting the appropriate parameters for the ANN, and the uncertainty associated with the outcome of iterative, error backpropagation for optimization. These practical challenges are well-known and may prevent ANNs from realizing their full theoretical potential.

There were a few situations (e.g., the estimate of J_{TBCL} for the speech-relevant data) where the ANN did perform better than LWR. Indeed, the ANN displayed a comparable mean normalized RMSE in estimating the direct kinematics on the speech-relevant development data (i.e.,during the parameter optimization process, see Figure 4). This accuracy did not generalize to the test set, however.

The ANN appears to struggle with modeling lip protrusion (PRO), which represents the simplest articulator-task relationship in the model. PRO is only dependent on the articulator variable LX, and the relationship is linear. This may be due to the built-in, static nonlinearities in the ANN, which are biased away from learning perfectly linear relationships. By that same token, the ANN does well on the highly nonlinear tasks related to the tongue body and tongue tip. LWR, on the other hand, performs well for all tasks, in spite of its assumptions of linearity.

In most cases, estimating the kinematic relationships from the speech-relevant data is more accurate and more consistent across tasks. Differences are particularly dramatic for the direct kinematic estimates. The overall higher accuracies in estimating from the speech-relevant data may be due to data sparsity issues. The speech-relevant data do not fill the entire articulator space, but rather are confined to certain regions of that space. Consequently, a smaller amount of data is needed to ensure dense coverage of the relevant space and accurate modeling. To increase the data density of uniformly-distributed data, an enormous number of data points are needed since the total number increases exponentially with the points along any single dimension. These trends are encouraging for future work on characterizing the kinematics of speech production, as opposed to vocal tract kinematics more generally. Attempts to estimate kinematic relationships from real speech-relevant data will require less data overall.

Finally, we observe that LWR outperforms ANN regardless of the quantity of training data. Moreover, LWR is very insensitive to the amount of training data, unlike the ANN which suffers from very high errors when the amount of training data is small. We also note that the performance of both algorithms seems to plateau around 3000 data points, indicating that the 5000-point data sets used for full evaluation were of sufficient size to provide stable results. This result also implies that very modest amounts of data are sufficient for learning the kinematic relationships with these methods, especially LWR.

Results from our preliminary experiments on real speech data were promising. Although observed estimation errors were consistently higher than errors on synthetic data, they were near the spatial resolution of our rtMRI data. It is difficult to interpret the specific implications of these errors because it depends on how these estimates get used downstream (e.g., features for an automatic speech recognizer (Ghosh and Narayanan, 2011)). Still, we can identify several sources of error that cause the discrepancy in errors between synthetic and real speech data. First, the simplified geometry and relatively fewer degrees of freedom in CASY/TADA versus a real vocal tract artificially reduce the error

on synthesized data, which is generally to be expected when using synthetic data. Errors may also reflect limitations in the acquisition and processing of rtMRI data. Those data are noisy, and the spatial and temporal resolution give only a limited view of the phenomena in question. Moreover, extraction of articulatory and task parameters (such as TTCD) from these data introduces noise through measurement uncertainty and approximation errors. Finally, our estimation techniques can likely be refined in the way they handle the variability and uncertainty in measurements and representations being used. Ongoing and future work will be aimed at improving upon these limitations, in particular the estimation techniques and methods for fitting them appropriately.

5. Conclusion

We have described and evaluated two statistical models which can estimate the direct and the differential kinematics of a complex, nonlinear motor system from data. Both Artificial Neural Networks and Locally-Weighted Regression were trained, optimized and tested on several data sets of synthesized speech production data. Accuracies were appear high enough to facilitate further use of these methods for estimating the kinematic relationships of real speech production data.

Although synthesized data was used to facilitate evaluation of differential kinematic estimation accuracy, we have taken care to develop a methodology which is entirely repeatable on real speech production data in future extensions of this work. The estimation and evaluation of differential kinematics is a crucial aspect of this work, and the potential applications for these kinematic relationships for characterization of the speech production system was reviewed.

It was observed that the accuracies of both statistical methods were high, even with a relatively modest amount of data. The best accuracies resulted from LWR. Moreover, LWR appears relatively insensitive to the amount of training data available. LWR also has many desirable qualities from a practical standpoint, such as few free parameters and a training procedure with an analytical solution. The assumptions of local linearity might be viewed as limiting, but are seen here as entirely appropriate in context of similar assumptions that are widely made on motor control formulations and applications. Indeed, these assumptions did not seem to be any hindrance in the experiments presented here. Still, it must be noted that the performance of ANNs was very close to LWR in many aspects of the evaluation. ANNs are very powerful in theory, but may suffer from the practical difficulties associated with iterative training procedures like error backpropagation.

Technical challenges still remain in terms of improving estimation accuracy even further. Alternative estimation techniques could be employed to that end. One possibility would be the Bayesian formulation of LWR developed by Ting et al. (2008). Based on a probabilistic formulation of regression, this technique allows for automatic optimization of the locality parameters. More advanced Neural Network architectures, such as Deep Belief Networks (Hinton et al., 2006), are also promising candidates.

Additional challenges remain in assessing whether the differential kinematics need to be estimated to the level of detail espoused here, or whether further approximations are appropriate for speech. For instance, it is not entirely clear to what degree the Jacobian is actually pose-dependent, even though the mathematics presented here express it as such. Given the nature of the expected nonlinearities, we suspect that the Jacobian will indeed be quite dependent on the pose. This claim should be assessed empirically, however, by inspecting the pose-by-pose changes to Jacobians estimated from real speech data.

The ultimate goal of this work is to utilize knowledge of kinematic relationships in order to gain insight into interesting, longstanding and unaddressed problems in the study of speech production. As such, a key extensions of this work will be to apply these estimation methods to real articulatory and acoustic data. The ability to estimate the differential kinematics, in particular, of real speech production data will provide insight and facilitate characterization of the speech production system. Notable applications will include the ability to ascertain the nature of speech production goals and to gain insight into acoustic-to-articulatory inversion.

It is important to note that the utility of these methods depends heavily on obtaining high-quality speech production data. The recently collected MRI-TIMIT database (Narayanan et al., 2011) may provide a useful platform for many applications of these methods. However, the addition of muscle activation data for speech would enhance the possibilities even further, as would the acquisition of 3-dimensional data at high frame rates and the ability to gather clean audio from MRI.

6. Acknowledgments

This work was supported by NIH NIDCD grant 02717, NIH R01 grant DC008780, NIH grant DC007124, as well as a graduate fellowship from the

Annenberg Foundation. We would also like to acknowledge Elliot Saltzman for his technical insights, and Hosung Nam for his help with understanding TADA.

Appendix A. CASY Equations

The articulatory variables are denoted q_i and the task variables are denoted by x_j. Constants include $l_{ut} = 1.1438$, $a_{ut} = -0.1888$, $l_{lt} = 1.1286$, $o_x = 0.7339$, $o_y = -0.4562$, $r_{ts} = 0.4$, $r_{tb} = 0.02$, $a_{tc} = 1.7279$, $l_{tb} = 0.8482$ and $s_{tb} = 4.48$.

The equations for calculating the lip-related tasks from the articulatory variables are:

$$x_{PRO} = q_{lx} \tag{A.1}$$

$$x_{LA} = l_{ut} sin(a_{ut}) + l_{lt} cos(q_{ja}) + q_{uy} - q_{ly} \tag{A.2}$$

The equations for calculating the tongue body tasks from the articulatory variables are:

$$a = q_{cl} sin(q_{ja} + q_{ca}) \tag{A.3}$$

$$b = -q_{cl} cos(q_{ja} + q_{ca}) \tag{A.4}$$

$$x_{TBCL} = cos^{-1} \frac{a - o_x}{\sqrt{(a - o_x)^2 + (b - o_y)^2}} \tag{A.5}$$

$$x_{TBCD} = r_{ts} - \sqrt{(a - o_x)^2 + (b - o_y)^2} + r_{tb} \tag{A.6}$$

The equations for calculating the tongue-tip-related tasks from the articulatory variables are:

$$c = q_{ja} + q_{ta} + s_{tb}(q_{cl} - l_{tb}) \tag{A.7}$$

$$d = a + r_{tb} sin(q_{ja} + a_{tc}) + q_{tl} sin(c) \tag{A.8}$$

$$e = b - r_{tb} cos(q_{ja} + a_{tc}) - q_{tl} cos(c) \tag{A.9}$$

$$x_{TTCL} = cos^{-1} \frac{d - o_x}{\sqrt{(d - o_x)^2 + (e - o_y)^2}} \tag{A.10}$$

$$x_{TTCD} = r_{tb} - \sqrt{(d - o_x)^2 + (e - o_y)^2} \tag{A.11}$$

Appendix B. Stimuli

We synthesized the following sentences using CASY/TADA to create the speech-relevant data set, including the articulator vectors and their accompanying task vectors. These 30 sentences represent the first 30 sentences of the MOCHA-TIMIT corpus, as developed by Wrench and Hardcastle (2000).

1. This was easy for us.
2. Is this seesaw safe?
3. Those thieves stole thirty jewels.
4. Jane may earn more money by working hard.
5. She is thinner than I am.
6. Bright sunshine shimmers on the ocean.
7. Nothing is as offensive as innocence.
8. Why yell or worry over silly items?
9. Where were you while we were away?
10. Are your grades higher or lower than Nancy's?
11. He will allow a rare lie.
12. Will Robin wear a yellow lily?
13. Swing your arm as high as you can.
14. Before Thursday's exam, review every formula.
15. The museum hires musicians every evening.
16. A roll of wire lay near the wall.
17. Carl lives in a lively home.
18. Alimony harms a divorced man's wealth.
19. Aluminium cutlery can often be flimsy.
20. She wore warm, fleecy, woolen overalls.
21. Alfalfa is healthy for you.
22. When all else fails, use force.
23. Those musicians harmonize marvellously.
24. Although always alone, we survive.
25. Only lawyers love millionaires.
26. Most young rabbits rise early every morning.
27. Did dad do academic bidding?
28. Beg that guard for one gallon of petrol.
29. Help Greg to pick a peck of potatoes.
30. Get a calico cat to keep the rodents away.

References

J.H. Abbs and V.L. Gracco. Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology*, 51:705–723, 1984.

S. Al Moubayed and G. Ananthakrishnan. Acoustic-to-articulatory inversion based on local regression. In *Proceedings of INTERSPEECH*, pages 937–940, 2010.

G. Ananthakrishnan, D. Neiberg, and O. Engwall. In search of non-uniqueness in the acoustic-to-articulatory mapping. In *Proceedings of INTERSPEECH*, pages 2799–2802, 2009.

B.S. Atal and O. Rioul. Neural networks for estimating articulatory positions from speech. *Journal of the Acoustical Society of America*, 86, 1989.

B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63:1535–1555, 1978.

C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *AI Review*, 11:11–73, 1997.

G. Bailly, R. Laboissière, and J. L. Schwartz. Formant trajectories as audible gestures: an alternative to speech synthesis. *Journal of Phonetics*, 19:9–23, 1991.

A. Balestrino, G. De Maria, and L. Sciavicco. Robust control of robotic manipulators. In *Proceedings of the 9th IFAC World Congress*, volume 5, pages 2435–2440, 1984.

D.J. Bennet and J.M. Hollerbach. Autonomous calibration of single-loop closed kinematic chains formed by manipulators with passive end-point constraints. *IEEE Transactions on Robotic Automation*, 7:597–606, 1991.

N.A. Bernstein. *The coordination and regulation of movements*. Pergamon Press, 1967. originally published in 1947.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

L.-J. Boë and G. Bailly. The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1):27–38, 1992.

E. Bresch and S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time mri. *IEEE Transactions in Medical Imaging*, 28(3):323–338, 2009.

D. Bullock, S. Grossberg, and F.H. Guenther. A self-organizing neural network model for redundant sensory-motor control, motor equivalence, and tool use. *Journal of Cognitive Neuroscience*, 5:408–435, 1993.

W. S. Cleveland. Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*, 74: 829–836, 1979.

W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local tting. *Journal of the American Statistical Association*, 83:596–610, 1988.

W. S. Cleveland, S. J. Devlin, and E. Grosse. Regression by local tting: Methods, properties, and computational algorithms. *Journal of Econometrics*, 37:87–114, 1988.

T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 23(6): 681–685, 2001.

G. Cybenko. Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

A. D'Souza, S. Vijayakumar, and S. Schaal. Learning inverse kinematics. In *Proceedings of CIRAS*, 2001.

W. Duch and N.J. Jankowski. Survey of neural transfer functions. *Neural Computing Surveys*, 2:163–212, 1999.

S. Fels, F. Vogt, K. van den Doel, J. Lloyd, and O. Guenter. Artisynth: Towards realizing an extensible, portable 3d articulatory speech synthesizer. In *Proceedings of the International Workshop on Auditory Visual Speech Processing*, pages 119–124, July 2005.

J.M. Gérard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan. A 3d dynamical biomechanical tongue model to study speech motor control. *Recent Research Development in Biomechanics*, 1:49–64, 2003.

J.M. Gérard, P. Perrier, and Y. Payan. 3d biomechanical tongue modeling to study speech production. In J. Harrington and M. Tabain, editors, *Speech Production: Models, Phonetic Processes, and Techniques*, pages 85–102. Psychology Press, 2006.

P. Ghosh and S. Narayanan. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *J. Acoust. Soc. Am. Express Letters*, 130(4):EL251–EL257, 2011.

P.K. Ghosh and S. Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 128(4):2162–2172, 2010.

F. Guenther. A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, 72:43–53, 1994.

F. Guenther. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102:594–621, 1995.

F. Guenther, M. Hampson, and D. Johnson. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105:611–633, 1998.

E. Guigon, P. Baraduc, and M. Desmurget. Computational motor control: Redundancy and invariance. *Journal of Neurophysiology*, 97 (1):331–347, 2007.

M. T. Hagan and M. Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5 (6):989–993, 1994.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

S. Hiroya and M. Honda. Determination of articulatory movements from speech acoustics using an hmm-based speech production model. In *Proceedings of ICASSP*, pages 437–440, 2002a.

S. Hiroya and M. Honda. Acoustic-to-articulatory inverse mapping using an hmm-based speech production model. In *Proceedings of ICSLP*, pages 2305–2308, 2002b.

S. Hiroya and M. Honda. Speech inversion for arbitrary speaker using a stochastic speech production model. In *Proceedings of the Interdisciplinary Workshop on Speech Dynamics by Ear, Eye, Mouth and Machine*, pages 9–14, 2003.

S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an hmm-based speech production model. *IEEE Trans Speech and Audio Processing*, 12(2):175–185, 2004.

J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. Accurate recovery of articulator positions from acoustics: new conclusions based on human data. *Journal of the Acoustical Society of America*, 100:1819–1834, 1996.

J.M. Hollerbach. Computers, brains and the control of movement. *Trends in Neurosciences*, 5:189–192, 1982.

J.M. Hollerbach and C.W. Wampler. The calibration index and taxonomy of robot kinematic calibration methods. *The International Journal of Robotics Research*, 15:573, 1996.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal function approximators. *Neural Networks*, 2, 1989.

K. Iskarous, L. Goldstein, D.H. Whalen, M. Tiede, and P. Rubin. Casy: The haskins configurable articulatory synthesizer. In *Proceedings of ICPhS*, 2003.

Wu. J.-M. Multilayer potts perceptrons with levenberg-marquardt learning. *IEEE Transactions on Neural Networks*, 19(12):2032–2043, 2008.

R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

M. Jordan. Constrained supervised learning. *Journal of Mathematical Psychology*, 36:396–425, 1992.

M. Jordan and D. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

M.I. Jordan and R.A. Jacobs. Modular and hierarchical learning systems. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

T. Kaburagi and M. Honda. Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database. In *Proceedings of ICSLP*, 1998.

C. Kello and D.C. Plaut. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *Journal of the Acoustical Society of America*, 116(4): 2354–2364, 2004.

S. Kelso, B. Tuller, E. Vatikiotis-Bateson, and C. Fowler. Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology*, 10(6):812–832, 1984.

O. Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal of Robotics and Automation*, 3(1):43–53, 1987.

A. Lammert, L. Goldstein, and K. Iskarous. Locally-weighted regression for estimating the forward kinematics of a geometric vocal tract model. In *Proceedings of INTERSPEECH*, 2010.

A. Lammert, V. Ramanarayanan, L. Goldstein, K. Iskarous, E. Saltzman, H. Nam, and S. Narayanan. Statistical estimation of speech kinematics from real-time mri data. *Journal of the Acoustical Society of America*, 130(4):2549, 2011.

A.C. Lammert, D.P.W. Ellis, and P. Divenyi. Data-driven articulatory inversion incorporating articulatory priors. In *Proceedings of SAPA*, pages 29–34, 2008.

S. Lawrence, A.C. Tsoi, and A.D. Black. Function approximation with neural networks and local methods: Bias, variance and smoothness. In *Proceedings of Australian Conference on Neural Networks*, pages 16–21, 1996.

R.S. McGowan and M.A. Berger. Acoustic-articulatory mapping in vowels by locally-weighted regression. *Speech Communication*, 126(4):2011–2032, 2009.

P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

P. Mermelstein and M. Schroeder. Determination of smoothed cross-sectional area functions of the vocal tract from formant frequencies. *Journal of the Acoustical Society of America*, 37:1186, 1965.

V. Mitra, Y. Ozbek, H. Nam, X. Zhou, and C.Y. Espy-Wilson. From acoustics to vocal tract time functions. In *Proceedings of ICASSP*, 2009.

V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein. Retrieving tract variables from acoustics: a comparison of different machine learning strategies. *IEEE Journal of Selected Topics on Signal Processing*, 4:1027–1045, 2010. Sp. Iss. on Statistical Learning Methods for Speech and Language Processing.

V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein. Tract variables for noise robust speech recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 2011. To Appear.

B.W. Mooring, Z.S. Roth, and M.R. Driels. *Fundamentals of Manipulator Calibration*. Wiley Interscience, 1991.

D. Mottet, Y. Guiard, T. Ferrand, and R. Bootsma. Two-handed performance of a rhythmical fitts task by individuals and dyads. *Experimental Psychology: Human Perception and Performance*, 27: 1275–1286, 2001.

K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda. On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum. In *Proceedings of ICASSP*, 2006.

Y. Nakamura and H. Hanafusa. Inverse kinematics solutions with singularity robustness for robot manipulator control. *Journal of Dynamic Systems, Measurement, and Control*, 108:163–171, 1986.

J. Nakanishi, M. Mistry, J. Peters, and S. Schaal. Operational space control: A theoretical and empirical comparison. *International Journal of Robotics Research*, 27(6):737–757, 2008.

H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. Tada: An enhanced, portable task dynamics model in matlab. *Journal of the Acoustical Society of America*, 115(5):2430–2430, 2004.

H. Nam, L. Goldstein, C. Browman, P. Rubin, M. Proctor, and E. Saltzman. *TADA (TAsk Dynamics Application) manual*, 2006.

H. Nam, V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C.Y. Espy-Wilson, and M. Hasegawa-Johnson. A procedure for estimating gestural scores from natural speech. In *Proceedings of ICASSP*, 2010.

S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, 109:2446, 2004.

S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu. A multimodal real-time mri articulatory corpus for speech research. In *Proceedings of INTERSPEECH*, 2011.

S. Panchapagesan and A. Alwan. A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model. *Journal of the Acoustical Society of America*, 129(4):2144–2162, 2011.

G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network train on x-ray microbeam data. *Journal of the Acoustical Society of America*, 92:688–700, 1992.

Y. Payan and P. Perrier. Synthesis of v–v sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Communication*, 22:187–205, 1997.

P. Perrier, H. Lœvenbruck, and Y. Payan. Control of tongue movements in speech: The equilibrium point hypothesis perspective. *Journal of Phonetics*, 24:53–75, 1996.

P. Perrier, Y. Payan, M. Zandipour, and J. Perkell. Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *Journal of the Acoustical Society of America*, 114(3):1582–1599, 2003.

C. Qin and M.Á. Cerreira-Perpiñán. An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. In *Proceedings of INTERSPEECH*, 2007.

C. Qin and M.Á. Cerreira-Perpiñán. Articulatory inversion of american english /r/ by conditional density modes. In *Proceedings of INTERSPEECH*, 2010.

M.G. Rahim, W.B. Kleijn, J. Schroeter, and C.C. Goodyear. Acoustic-to-articulatory parameter mapping using an assembly of neural networks. In *Proceedings of ICASSP*, pages 485–488, 1991.

K. Richmond. A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Proceedings of INTERSPEECH*, pages 577–580, 2010.

P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede, and C. Browman. Casy and extensions to the task-dynamic model. In *Proceedings of the 1st ETRW on Speech Production Modeling, Autrans, France*, 1996.

D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press, 1986a.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088), 1986b.

E. Saltzman and D. Byrd. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19(4):499–526, 2000.

E. Saltzman and J.A.S. Kelso. Skilled actions: A task dynamic approach. *Psychological Review*, 94:84–106, 1987.

E. Saltzman and K.G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.

E. Saltzman, M. Kubo, and C.C. Tsao. Controlled variables, the uncontrolled manifold, and the task-dynamic model of speech production. In Divenyi et al., editor, *Dynamics of Speech Production and Perception*. IOS Press, 2006.

J.P. Scholz and G. Schöner. The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental Brain Research*, 126:189–306, 1999.

J. Schroeter and M.M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech Audio Processing*, 2:133–150, 1994.

L. Sciavicco and B. Siciliano. *Modelling and Control of Robot Manipulators*. Springer, 2005.

Y. Shiga and S. King. Estimating detailed spectral envelopes using articulatory clustering. In *Proceedings of INTERSPEECH*, 2004.

M.E. Sklar. Geometric calibration of industrial manipulators by circle point analysis. In *Proceedings of the 2nd Conference on Recent Advances in Robotics*, pages 178–202, 1989.

J.F. Soechting. Does position sense at the elbow joint reflect a sense of elbow joint angle or one of limb orientation? *Brain Research*, 248:392–395, 1982.

J.A. Ting, A. D'Souza, S. Vijayakumar, and S. Schaal. A bayesian approach to empirical local linearization for robotics. In *Proceedings of ICRA, Pasadena, CA.*, 2008.

T. Toda, A.W. Black, and K. Tokuda. Mapping from articulatory

movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pages 31–36, 2004.

T. Toda, A.W. Black, and K. Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3), 2008.

A. Toledo, M. Pinzolas, J.J. Ibarrola, and G Lera. Improvements of the neighborhood based levenberg-marquardt algorithm by local adaptation of the learning coefficient. *IEEE Transactions on Neural Networks*, 16(4):988–992, 2005.

F. Vogt, O. Guenther, A. Hannam, K. van den Doel, J. Lloyd, L. Vilhan, R. Chander, J. Lam, C. Wilson, K. Tait, D. Derrick, I. Wilson, C. Jaeger, B. Gick, E. Vatikiotis-Bateson, and S. Fels. Artisynth designing a modular 3d articulatory speech synthesizer. *Journal of the Acoustical Society of America*, 117(4):2542, May 2005.

F. Vogt, J.E. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S.S. Fels. An efficient biomechanical tongue model for speech research. In *Proceedings of ISSP 06*, pages 51–58, 2006.

H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, 21(5):417–427, 1973.

C.W. Wampler. Manipulator inverse kinematic solutions based on vector formulations and damped least squares methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 16:93–101, 1986.

D.E. Whitney. Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man-Machine Systems*, 10: 47–53, 1969.

B.M. Wilamowski, N.J. Cotton, O. Kaynak, and G. Dündar. Computing gradient vector and jacobian matrix in arbitrarily connected neural networks. *IEEE Trans on Industrial Electronics*, 55(10): 3784–3790, 2008.

R. Winkler, S. Fuchs, P. Perrier, and M. Tiede. Biomechanical tongue models: An approach to studying inter-speaker variability. In *Proceedings of INTERSPEECH*, 2011a.

R. Winkler, L. Ma, and P. Perrier. A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing. In *Proceedings of ISSP*, 2011b.

W.A. Wolovich and H. Elliot. A computational technique for inverse kinematics. In *Proceedings of the 23rd IEEE Conference on Decision and Control*, pages 1359–1363, 1984.

A. Wrench and W. Hardcastle. A multichannel articulatory speech database and its application for automatic speech recognition. In *Proceedings of the 5th Seminar on Speech Production*, pages 305–308, 2000.