Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 163–176

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

An adaptive neural control scheme for articulatory synthesis of CV sequences☆

Guangpu Huang, Meng Joo Er*

Computer Vision Laboratory, Nanyang Technological University, 50 Nanyang Avenue, 639798, Republic of Singapore

Received 1 September 2012; received in revised form 4 April 2013; accepted 13 April 2013
Available online 21 April 2013

Abstract

Reproducing the smooth vocal tract trajectories is critical for high quality articulatory speech synthesis. This paper presents an adaptive neural control scheme for such a task using fuzzy logic and neural networks. The control scheme estimates motor commands from trajectories of flesh-points on selected articulators. These motor commands are then used to reproduce the trajectories of the underlying articulators in a 2nd order dynamical system. Initial experiments show that the control scheme is able to manipulate the mass-spring based elastic tract walls in a 2-dimensional articulatory synthesizer and to realize efficient speech motor control. The proposed controller achieves high accuracy during on-line tracking of the lips, the tongue, and the jaw in the simulation of consonant–vowel sequences. It also offers salient features such as generality and adaptability for future developments of control models in articulatory synthesis.

Keywords: Articulatory synthesis; Speech motor control; Neural networks; Fuzzy logic; Mass spring damper

1. Introduction

There are mainly two types of synthesis methods in text-to-speech (TTS) applications: concatenative and articulatory synthesis (Birkholz, 2005). Concatenative synthesis uses the stored speech waveforms of phonemes or words pronounced by the human speakers to generate intelligible output. Its applications are limited to the languages and speakers available. In contrast, articulatory synthesis simulates the movements in the speech apparatus of human speakers for the TTS application. It has a stronger physiological basis and is able to produce a larger number of utterances than the concatenative method. In fact, the method offers additional benefits beyond TTS, in applications such as the facial animation (Badin et al., 2002), the medical treatment of speech disorders (Kröger et al., 2008), and the articulatory-phonetic studies in automatic speech recognition (King et al., 2007).

However, it remains a challenging task to reproduce the vocal tract trajectories through automatic control in current articulatory synthesis research. A complete articulatory synthesizer usually includes three functional components: an anatomical model, an acoustic model, and a control model. Studies on the anatomical and the acoustic models have developed rapidly in the past decades (Buchaillard et al., 2009; Birkholz et al., 2007; Cook, 1990), but there

☆ This paper has been recommended for acceptance by Prof. R.K. Moore.

* Corresponding author. Tel.: +65 9299 0661; fax: +65 6896 8757.
E-mail addresses: hu0002pu@e.ntu.edu.sg (G. Huang), emjer@ntu.edu.sg (M.J. Er).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.04.004

is a lack of study in the control model. The desired control model should be able to reproduce realistic articulatory trajectories in different phonetic contexts and with different speaking rate. Existing control models often operate manually in a codebook fashion, which applies a set of linguistic rules to define the articulatory targets such as the velocity and the position profile of a particular speech sound (e.g., a phone). Such a synthesis-by-rule approach was initially implemented in the cord-tract model of Ishizaka et al. (1975) and the task-dynamic articulatory model of Saltzman and Munhall (1989). In their approach, each phone usually has one spatial target in the codebook. The articulatory movements for the sequential phonetic strings such as syllables, words, and sentences, are generated by interpolating and/or approximating the targets (Birkholz et al., 2011; Perrier et al., 2005). Different from the codebook approach, Nelson (1983) suggested that the articulatory movements were the result of optimized control similar to that of a second-order dynamical system. Löfqvist and Gracco (2002) supported the view, and they observed that a cost minimization principle could well explain the trajectory curvature of the articulatory kinematics.

The main difficulty lies in the dynamics of speech motor control. In the literature, many methods have been proposed to model the the positions and velocities of the speech articulators during speech production, for example, the task dynamic model of Saltzman and Munhall (1989), the models of Perrier et al. (2003) and Buchaillard et al. (2009) based on the Equilibrium Point Hypothesis (EPH) of Feldman (1986). However, the speech dynamics are highly non-linear and contain many uncertainties, which are difficult to describe using the precise mathematical models. There is an urgent need for more adequate modeling methods to realize efficient speech motor control. In this paper, we re-formulated the articulatory dynamics using a mass-spring damper (MSD) in a 2-dimensional articulatory synthesizer. We used the fuzzy neural networks (FNNs) to deal with the unmodeled uncertainties and non-linearity in an adaptive neural controller. In contrast to using the fixed-structured neural networks for the non-linear modeling (Richmond, 2009), the proposed controller embedded a learning algorithm and an adaptive control law to determine the structure and the parameters of the neural topology simultaneously during off-line learning. In other words, FNNs learn the speech dynamics, or the mapping between the input and the output, and store the information in the neural topology. During on-line tracking, the controller estimated a series of motor commands from trajectories of flesh-points on the selected articulators. Then it used these motor commands to reproduce the trajectories for the underlying articulators. We used a set of consonant–vowel (CV) sequences to demonstrate the learning and the tracking process in the simulations. Then the controller manipulated the MSD to reproduce the smooth trajectories in the selected articulators. Our experiments showed that it achieved high accuracy during on-line tracking of the lips, the tongue, and the jaw in the CV sequences.

The rest of the paper is organized as follows. Section 2 introduces the background and the theoretical basis of speech motor control. Section 3 formulates the articulatory dynamics in the MSD based 2-D vocal tract system. Section 4 describes the structure, the learning algorithm, and the adaptive laws of the proposed E-FNN controller. Section 5 describes the experimental settings and the simulation procedures. Section 6 discusses the results. Section 7 concludes this paper.

2. Overview of speech motor control

In speech motor control, the equation of motion governs the dynamics of the articulators. It is analogous to the MSD (Kröger et al., 1995; Perrier and Ostry, 1996; Kelso et al., 1986), which follows Newton's law

$$u + F_e + F_f = M\ddot{z} + B\dot{z} + Kz, \tag{1}$$

where M, B, and K are the mass, damping, and stiffness coefficients of the speech articulators (e.g., the tongue tip and the lower lip) in the anatomical model of an articulatory synthesizer. F_e is the external force due to the gravity factor and the air pressure inside the tract, and F_f is the friction force between the adjacent muscular structures, which can be assumed to be negligible due to the saliva. We describe the articulators using the motion vectors, termed vocal tract variables (TVs), where z, $\dot{z}$, and $\ddot{z}$ are the position, the velocity, and the acceleration parameters, respectively. u is the input force or the activation level of the muscular structures which control the TVs. We refer to the set of muscular activation forces as the motor variables (MVs).

In comparison, the task dynamic model of Saltzman and Munhall (1989) distinguishes three kinds of variables at two levels: the activation coordinates at the inter-gestural level, and the model articulator coordinates (the actual spatial position of the articulators) and tract-variable coordinates (the location and degree of constriction of the articulators) at the inter-articulator level. In this paper, the MVs are analogous to the activation coordinates. The TVs are not replicates of Saltzman and Munhall's model articulator coordinates, nor are they the same as the tract-variable coordinates.

Instead they are the pellet position/coordinates at the selected constriction locations on the articulators (more on these variables in Section 3).

The equation of motion describes the quasi-incompressibility of the speech articulators during the speech production (Kim and Gomi, 2007). It yields close-loop solutions by choosing the appropriate time-variant variables. For example, the Equilibrium Point Hypothesis (EPH) of Feldman (1986) used the equilibrium positions as the time-variant variables, the shift of which results in the movements of the articulators. Perrier and Ostry (1996), Perrier et al. (2003), and Buchaillard et al. (2009) applied the EPH control concept in a finite element model of the tongue, and solved the differential equation using combined Newton-Raphson and Newmark method. In contrast, Saltzman and Munhall (1989) considered the stiffness coefficients as the time-varying variables. They used a pseudo-Jacobian inversion matrix to calculate the gestural control parameters for the desired articulatory movements in the differential equation. The concept still resembles the EPH method, since the stiffness directly affects the velocity with which the equilibrium length is restored (Boersma, 1998). The concept is also used in the articulatory synthesizer of Birkholz (2005), Kröger et al. (1995, 2009). However, it requires explicitly defining the gestural scores and the activation intervals for the system state profile $[\dot{z}, z]$ in the control model, which are highly error prone especially at the phonetic boundaries (Kelso et al., 1986). Another way to solve the differential equation is to use the time-varying input force functions to reproduce the system state profile $[\dot{z}, z]$, the velocity and position trajectories (Kröger et al., 1995). The coefficients M, K, and B assume values that are close to human tissues in the vocal tract. In this manner, the equation of motion in (1) simplifies to an ordinary differential equation. For example, van den Doel and Ascher (2008) formulated a wall displacement model

$$p(x, t) = M\ddot{z}(x, t) + B\dot{z}(x, t) + K\vec{z}(x, t), \tag{2}$$

where the driving force is from the air pressure p inside the tube. Additional discretization techniques such as the leap-frog scheme (Boersma, 1998) and the Newmark methods (van den Doel and Ascher, 2008) are then used to solve the equation during articulatory and acoustic simulation. One major drawback of this approach is the high computation cost which renders it inefficient for on-line articulatory control.

Moreover, the dynamic MSD system in (1) is highly non-linear and contains uncertainties which are difficult to describe using precise mathematical model. There are mainly three difficulties.

1. Human vocal system consists of soft tissues as well as bony structures, e.g., the hard palate. Consequently, the MSD system contains unmodeled variabilities in the M, B, and K parameters, which vary from speaker to speaker (e.g., physiological differences) and for the same speaker under different conditions (e.g., emotional states). The stiffness of muscular tissues also changes during activation (Duck, 1990; Perrier et al., 2003).
2. The articulatory movements are affected by the phonetic structure of continuous speech, which introduces unmodeled variabilities.
3. During speech production, the modeling of constriction is not linear. Though the articulatory movements are smooth between vowel targets, the transitions to/from the consonants such as plosives, nasals, and laterals, are not so. For example, when the tongue tip hits the alveolar ridge during [d/t] production, the collision introduces points of discontinuity in the vocal tract at the onset of the closure, rendering the model non-linear (Birkholz et al., 2011).

Therefore, there is an urgent need for more adequate dynamic modeling methods to deal with the above unmodeled variabilities and non-linearity to realize efficient speech motor control.

Neural networks (NNs) have shown advantages in non-linear modeling of dynamic control systems. For example, Saltzman and Munhall (1989) proposed to use Jordan's recurrent neural networks (RNNs) (1986) to incorporate the temporal dynamics and learning algorithm in the control model. Hirayama et al. (1993) applied NNs to learn the inverse dynamics of speech motor control. More recently Fang (2009) used a general regression neural model to infer motor commands from the articulatory measurements. However, these are fix-structured NNs, which use a trial-by-error approach to determine the parameter and structure in the neural controller. As a result, the controller performance is subject to the experimenter's decision rather than the property of the dynamic system. In this aspect, NNs with fuzzy logic, or FNNs are more appropriate than the fix-structured NNs (Wang, 1997). They have been used to improve speech motor control in articulatory synthesis. For example, Kröger et al. (2009) used self-organizing maps to learn the motor commands and the tract variables from phonetic sequences, and obtained encouraging results in the articulatory synthesizer. FNNs have yet to reach their full potential.

Previously we have introduced an adaptive neural controller, termed the generalized dynamic fuzzy neural network (GD-FNN) controller (Wu et al., 2001; Er and Gao, 2003). The controller has shown excellent performance in terms of tracking accuracy and computational efficiency for several non-linear dynamic systems with unmodeled variabilities, e.g., an inverted pendulum, a robot manipulator (Gao and Er, 2003), and a drug delivery system (Gao and Er, 2005). In this study, we applied the adaptive neural control model to reproduce the articulatory trajectories of the vocal apparatus in a 2-dimensional (2-D) articulatory synthesizer. The GD-FNN infers knowledge about the articulatory dynamics and stores the information in the neural structures and the fuzzy logics. The proposed control scheme is an extended version of GD-FNN, referred to as E-FNN. It integrates the radial basis function neural network (RBF-NN), the fuzzy inference network (FIN), and the recurrent neural network (RNN) in one neural topology. The recurrent layer is added to the original GD-FNN to deal with the temporal dynamics in the articulatory speech patterns (Jordan, 1986). The complete E-FNN controller also embeds a learning algorithm and an adaptive control law to determine the structure and the parameters of the neural topology simultaneously. Our main hypothesis is that it is possible to deal with the uncertainty and the non-linearity in the mapping between the muscle activities, the MVs, and the articulatory trajectories, the TVs, in speech motor control. Unlike the TVs, the MVs are usually hard to measure or not completely retrievable in human speech production. In this study, the E-FNN learns to predict the MVs from the TVs using the generalization abilities of fuzzy logics and NNs. We then couple the E-FNN model with a proportional integral derivative (PID) controller to manipulate a MSD system to reproduce the continuous and smooth articulatory trajectories of the desired consonant–vowel (CV) sequences. We test the tracking accuracy of the E-FNN controller on electromagnetic articulography (EMA) data of the vocal tract in CV articulation.

3. Articulatory dynamics

The controllability canonical form for a 2nd order time-variant non-linear system is (Slotine and Li, 1991),

$$\ddot{z}(t_s) = f_n(\underline{z}, t_s) + g_n(\underline{z}, t_s)u(t_s) + d_n(t_s), \tag{3}$$

where $\underline{z} = [\dot{z}, z]$ is the state vector, velocity and position, of the system, f_n and g_n represent the non-linearities of the mapping from the input u to the output z, and d represents the uncertainties and external disturbances of the dynamic system, and d_n is the unmodeled uncertainties. If we further define the non-linear dynamic function $f_n(\underline{z}, t_s) = f(\underline{z}, t_s) + \Delta f(\underline{z}, t_s)$, and the control gain $g_n(\underline{z}, t_s) = g + \Delta g(\underline{z}, t_s)$, where f and g are the nominal parts, Δf and Δg are the unknown parts or the uncertainties of f and g (Lin and Li, 2012), the canonical form can be re-written as,

$$\ddot{z}(t_s) = f(\underline{z}, t_s) + gu(t_s) + d(t_s), \tag{4}$$

where $d(t_s) = \Delta f(\underline{z}, t_s) + \Delta g(\underline{z}u(t_s), t_s) + d_n(t_s)$ is the unknown uncertainties. From the equation of motion of the MSD system in (1) and the controllability canonical form in (4), we have

$$f = -\frac{B}{M}\dot{z}(t) - \frac{K}{M}z(t), \tag{5}$$

and

$$g = \frac{1}{M}. \tag{6}$$

Since $g \neq 0$ for all $\vec{z}$, the system is controllable (Lin and Li, 2012; Slotine and Li, 1991; Gao and Er, 2003). Furthermore, the function f and d are assumed to be bounded in human vocal system.

We focus on the control of the vocal tract including the lips, the tongue, and the jaw in a 2-D articulatory synthesizer, as shown in Fig. 1, which was constructed by Mermelstein (1973) based on the X-ray image of a human speaker. The vocal tract has elastic walls, which are analogous to the MSD. The TVs are the pellet points at the selected constriction locations on the articulators, as shown in Fig. 1. The 12 TVs include the x–y coordinates of the tongue root (TR_x, TR_y) relative from its neutral or resting position, the tongue body (TB_x, TB_y), the tongue tip (TT_x, TT_y), the lower lip (LL_x, LL_y), the upper lip (UL_x, UL_y), and the lower incisor (LI_x, LI_y). The MVs represent the muscular forces, which underly the TVs in the 2-D vocal tract. The 8 MVs cover one intrinsic tongue muscle: superior longitudinal (*SL*), which retracts or flaps the tongue tip; four extrinsic tongue muscles: anterior genioglossus (*GGa*), posterior genioglossus (*GGp*), hyoglossus (HG) and styloglossus (*SG*), which change the shape and position the tongue dorsum: body and root; three facial muscles: masseter (MA) which raises the lower jaw, risorius (*RO*) and orbicularisoris (*OO*) which

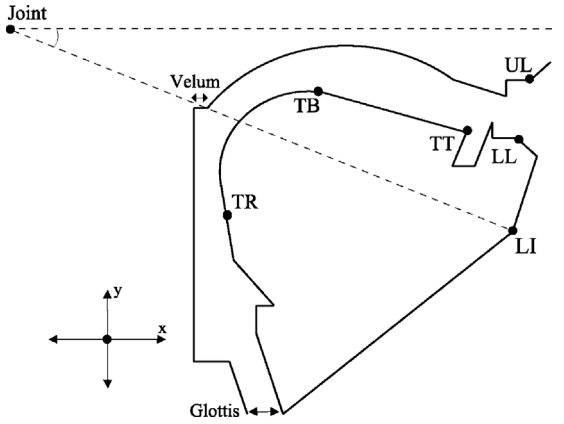

Fig. 1. Illustration of Mermelstein's 2-D articulatory mesh and the location of vocal tract variables.

constrict, round, and spread the lips. The vocal cords, shown as the glottis in Fig. 1, are not included in this model for two reasons. Firstly the control of vocal cords is more effective with stiffness parameters than the MV parameters (Flanagan et al., 1975). Secondly the vocal cords can cause non-unique mapping between the TVs and the MVs because it can compensate for vocal tract changes in speech production (Schroeter and Sondhi, 1994). For example, if we are to model the simple voicing contrast for the [p/b] and the [t/d] pairs, additional control variables regarding the timing of glottal excitation need to be specified in the vocal cords. Therefore, in the present model, it is not used as a control variable.

4. Neural control scheme

As shown in Fig. 2(a), during off-line training, the proposed E-FNN model learns the inverse characteristics between the input MV, u, and the output TV, z, in the dynamic MSD system. Since the exact MVs are unknown, and the parameters M, B, and K vary from speaker to speaker and for the same speaker in different phonetic contexts, the E-FNN controller uses the reference MVs, u_r, (details given in Section 5) and the desired TVs, z_d, to learn the system non-linearities

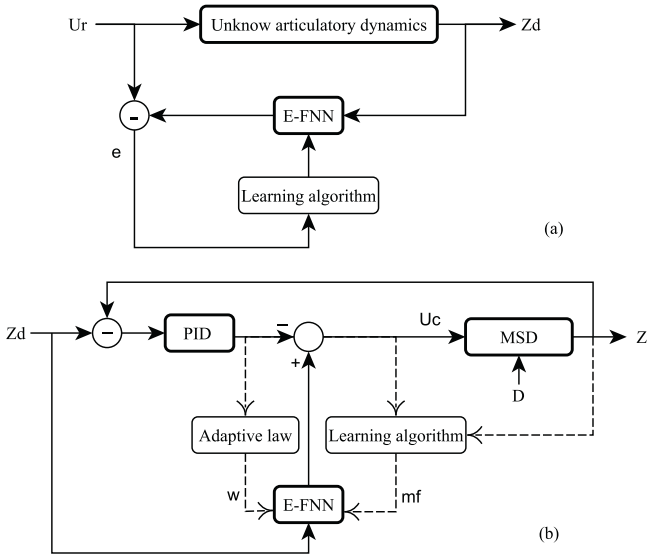

Fig. 2. Structure and data flow in the proposed fuzzy neural controller. (a) Off-line learning on the training data pairs (u_r, z_d). u_r is the reference motor variables, z_d is the desired tract variables, and e is the tracking error. (b) On-line tracking of the desired articulatory trajectories z_d in the MSD based vocal tract system. In the MSD block, u_c is the retrieved control signal, D represents the uncertainties, and z is the output of the dynamic system. In the E-FNN block, w is the weight parameter, and mf denotes the membership function (cf. Section 4.2).

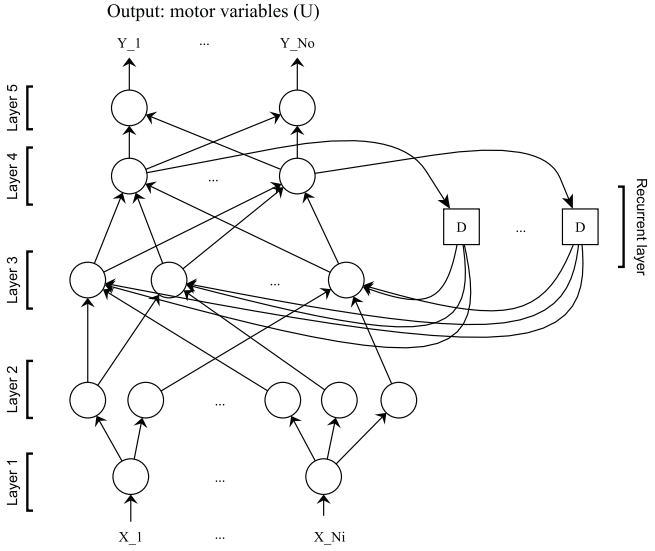

Fig. 3. Architecture of the extended fuzzy neural network.

and dynamics through an embedded learning algorithm. Using the training data pairs, the algorithm determines the structure and the parameters of the E-FNN such as the number of hidden neurons and the weights systematically and automatically through an iterative supervised learning (Section 4.2). During on-line tracking, as shown in Fig. 2(b), instead of looking for exact MVs, the PID controller generates the compensation output and the tracking error at each sample time for the overall control system. It embeds an adaptive control law, which uses the error rate as the weight update criteria and stores the system dynamics and the mapping functions in the E-FNN (Section 4.3). In this manner, the E-FNN controller infers the muscular activation patterns from trajectories of flesh-points on the selected articulators.

4.1. E-FNN structure

The E-FNN architecture is shown in Fig. 3, which has a total of five layers. It incorporates the Takagi–Suegeno–Kang-type fuzzy inference system, the RBF-NN, and the RNN in a connectionist structure, which is extended from the GD-FNN (Wu et al., 2001; Er and Gao, 2003; Gao and Er, 2005). We add the recurrent layer to account for the temporal dynamics in speech motor control (Jordan, 1986). Nodes and links in layer one and two act as a fuzzifier, while nodes and links in layer four act as a defuzzifier. We use $x_i^{(l)}$ to denote the ith input of a node in the lth layer, and $y_i^{(l)}$ to denote its corresponding output in layer l. The function of the node in each layer is given in the following.

- Layer 1: The input layer. Each node transmits the input variable to the next layer directly:

$$y_i^{(1)} = x_i^{(1)}, \quad i = 1, \ldots, N_i. \tag{7}$$

  For the inverse control model, $N_i = 36$, which includes the position, velocity and acceleration of the 12 TVs: $[\ddot{z}, \dot{z}, z]$.
- Layer 2: The membership function layer. It specifies the degree to which an input variable belongs to a fuzzy set using Gaussian membership function:

$$y_i^{(2)} = \exp - \frac{(x_i - c_{ij})^2}{\sigma_{ij}^2}, \tag{8}$$

  where c_{ij} and σ_{ij}, $i = 1, \ldots, N_i$, $j = 1, \ldots, N_j$, are the center and the width of the Gaussian function for the jth term in the ith input variable. These parameters are obtained in the learning procedure.

- Layer 3: The rule layer. The number of nodes indicates the number of fuzzy rules. The output of a rule node indicates the firing strength of its corresponding rule, defined as

$$y_j^{(3)} = y^{(6)} \prod_{i=1}^{N_i} x_i^{(3)}, \tag{9}$$

where $y^{(6)}$ is the output of the recurrent layer.

- Layer 4: The weight layer. The TSK-type fuzzy output weights are obtained in the structure learning procedure. The node output $y_k^{(4)}$: $k = 1, \ldots, N_k$ is the weighted sum of the incoming signals, which is a fuzzy "OR" operation:

$$y_k^{(4)} = \sum_{k=1}^{N_k} x_k^{(4)} = \sum_{j=1}^{N_j} y_j^{(3)} w_{jk}, \tag{10}$$

which integrates the fired rules on the same consequence neuron. The weight is:

$$w_{jk} = K_0 + \sum_{i=1}^{N_i} K_i x_i, \tag{11}$$

where K's are manually-set and real-valued parameters.

- Layer 5: The defuzzification layer. Each node in this layer corresponds to one output variable. The output function is defined as

$$y_o^{(5)} = \frac{\sum_{k=1}^{N_k} x_k^{(5)} w_{ok}}{\sum_{k=1}^{N_k} x_k^{(5)}}, \tag{12}$$

where $x_k^{(5)} = y_k^{(4)}$, w_{ok} is the link weight from the kth term in layer four to the oth output variable in layer five, $o = 1, \ldots, N_o$, and $N_o = N_k$. In the control model, $N_o = 8$, which is the number of the MVs in the dynamic MSD system. The neural function generates a output in [0, 1], which is the normalized activation level of the MVs.

- Recurrent layer: it calculates the firing strength of the recurrent variable $r_k = y_k^{(4)}$ to the rule layer. The number of recurrent nodes is the same at that of the output node in layer four. The node acts as a delay line to account for the contextual information in the temporal patterns. The node function is defined as

$$y_k^{(6)} = \frac{1}{1 + e^{-r_k}}. \tag{13}$$

The function can be interpreted as a global membership function, which "remembers" the history of discourse in the recurrent variables (Jordan, 1986). The recurrent outputs are fed back to the rules nodes in layer three, which stores the firing history of the fuzzy rules.

4.2. Learning algorithm

The learning algorithm enables simultaneous learning of the E-FNN structure and parameter, which was proposed and implemented in our previous studies (Wu et al., 2001; Gao and Er, 2003; Er and Gao, 2003). Structure learning determines the number of membership functions in layer two and the number of fuzzy logic rules in layer three. Parameter learning determines the Gaussian parameters in layer two and the link weights in layer four, i.e., the membership function $y_i^{(2)}(c_{ij}, \sigma_{ij})$, and the weight parameters w_{jk}. It uses the semi-closed fuzzy set for membership learning and the linear least square method for weight learning. Structure learning automatically creates or deletes fuzzy rules according to the system error and the error reduction ratio in the E-FNN controller. Learning repeats for each input and output data-pair. The parameters and the structure of the E-FNN are tuned automatically on the training data. Initially there are no fuzzy rules in layer three, and they are created or deleted automatically as the learning proceeds. Detailed mathematical descriptions of the learning algorithm, convergence analysis, and stability analysis of the FNN based controller in dynamic modeling are given in (Gao and Er, 2003).

4.3. Adaptive control law

After obtaining the initial value of the weight vector w_{ij} during the learning process, the E-FNN based controller embeds an adaptive control law to adjust the vector to compensate for the modeling errors in the learning algorithm (Gao and Er, 2005). In this study, the E-FNN controller is connected with a PID controller via adaptive control, as shown in Fig. 2(b). The PID controller serves as a feedback compensator which also stabilizes the inverse dynamic modeling (Gao and Er, 2005; Lin and Li, 2012). The adaptive control law is designed as follows,

$$u_c(t_s) = u_{\text{E-FNN}}(z_d, t_s) - u_{PID}(t_s). \tag{14}$$

The PID control output is given by,

$$u_{PID} = K_p e(t) + K_i \int e(t)dt + K_d \dot{e}(t), \tag{15}$$

where e is the tracking error: $e(t) = z_d - z$ between the desired target position and the displacement of the MSD. The matrix $K = [K_p, K_i, K_d]$ contains real numbers, and the proper choice of K affects the convergence speech of the tracking performance. The adaptive law adjusts the weight vectors in layer three and four of the E-FNN to minimize the square error E between the desired target position and the estimated position,

$$E(t_s) = \frac{1}{2} u_{PID}^2. \tag{16}$$

The discrete gradient method is used to minimize E. The adaptive law of the weight vector is derived as (Wu et al., 2001),

$$\Delta_W = -\eta \frac{\delta E}{\delta W} \tag{17}$$

$$= -\eta \frac{\delta E}{\delta u_{\text{S-FNN}}} \frac{\delta u_{\text{S-FNN}}}{\delta W} \tag{18}$$

$$= -\eta \frac{\delta \frac{1}{2}(u_c(t_s) - u_{\text{S-FNN}}(t_s))}{\delta u_{\text{S-FNN}}} \frac{\delta u_{\text{S-FNN}}}{\delta W} \tag{19}$$

$$= \eta u_{PID}(t_s) \phi(z_d, t_s) \tag{20}$$

where $\eta > 0$ is the learning rate.

5. Simulation

Simulation includes two stages: off-line learning and on-line tracking. In the first stage, the learning algorithm decides the initial weight parameter and the fuzzy rules of the E-FNN topology. It models the inverse dynamics between the motor commands and the tract variables. For this stage, we need to train the E-FNN on parallel MV and TV data. Ideally the training data consist of MVs and TVs measured on the human speech apparatus, such as the electromyographic (EMG) and the EMA recordings. The EMG recordings describe the level of muscular forces in the EPH model, while the EMA recordings describe the corresponding articulatory movements.

5.1. Data preparation

In this study, we used the CV sequences from the multichannel articulatory (MOCHA) database, which consists of two speakers: one male (MSAK0) and one female (FSEW0), each uttering 460 TIMIT sentences (Wrench, 1999). 807 CV syllables are available in the training data. Each CV sequence has a syllable initial plosive (with or without stress) for every combination of the vowels [ɑ, i, e, ɒ, u] and the plosives [p/b, t/d, k/g] in the pilot study. The EMA data in MOCHA records the movements of the articulators, or the 12 TVs. Similar to Mermelstein's 2-D model, the bridge of the nose and the upper incisor are taken as the reference point in the x–y coordinates (Browman and Goldstein, 1992). The trajectory vectors are z-normalized to have zero mean and unit variance, similar to Richmond (2009). The EMA data have at a sampling rate of 500 Hz.

Table 1
Motor variables in the vocal tract and their reference activation levels, u_r, in the plosive–vowel sequences.

	Tongue					Jaw	Lips	
	GGa	GGp	SG	HG	SL	MA	OO	RO
p/b	0	0	0	0.5	0	0	1	1
t/d	0	0	1	0	1	0	0	0
k/g	1	1	0	0	0	0	0	0
ɑ	1	0	0	1	0	0	0	0
i	0	1	0	0	0	1	0	1
ɒ	0	0	1	0	0	0	1	0
u	0	1	1	0	0	1	1	0
e	0	0	0	0	0	0.5	0	0

Reliable EMG data are usually difficult to obtain in articulatory studies, for example, through needle insertion (Baer et al., 1988). In their experiments, Baer et al. (1988) observed that the tongue muscles, GG_a, GG_p, *SG*, and *HG* have distinctive level of EMG activation for the cardinal vowels in [pVp] sequences. For example, a single threshold of EMG level can distinguish a front vowel from a back vowel, and each vowel group has consistent activation patterns. The claim was supported by Buchaillard et al. (2009) in the modeling of tongue muscles for cardinal vowel production. To this end, we use an alternative set of reference MVs derived from linguistic studies and the existing EMG recordings to initialize the learning process. As shown in Table 1, each phone represents a cognitive linguistic unit. It corresponds to the motor activity of the muscles, which are normalized to the [0, 1] interval to suit the NN input. For example, *GGa* is activated (=1) for the front vowel [ɑ], *GGp* for the high vowels [i] and [u], *SG* for the back vowels [ɒ] and [u], and *HG* for the low vowel [ɑ]. To reduce the variability of jaw positions, the MA activation is lower for the low vowel [ɑ] and [ɒ] than for the high vowels [i] and [u]. For the plosive pairs, *OO* and *RO* is activated for the labial [p/b], *SL* and *SG* for the alveolar [t/d], *GGa* and *GGp* for the velar [k/g].

During on-line tracking, the E-FNN controller retrieves the muscular activations in the CV sequences and reproduces the desired articulatory trajectories. We used a step function to simulate the muscle activation from the consonant to the vowel. The reference MVs in Table 1 are then updated by the PID compensator on the phonetic segments. The off-line learning in this study is analogous to the babbling stage in human speech acquisition while the on-line tracking corresponds to the imitation stage (Kröger et al., 2009; Bailly, 1997).

5.2. *Off-line training*

The E-FNN learning algorithm operates at a sampling rate of 100 Hz. The MOCHA training data are divided randomly into five sets, four of which are used for off-line training, the other one used for on-line tracking. The structure and parameter of the E-FNN are determined simultaneously on the four training sets of MV and TV data pairs. Using the learning algorithm, a total number of 23 fuzzy rules are created after training. Fig. 4 shows that the root mean square error (RMSE) converges after training on 250 samples.

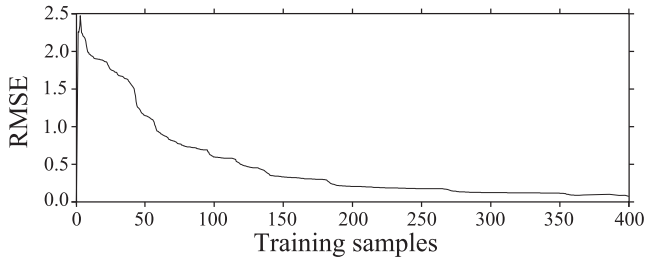

Fig. 4. Average RMSE rate of the fuzzy neural controller on MV inversion using the MOCHA training data.

Table 2
RMSE of the estimated articulatory trajectories in comparison with the EMA recordings.

	RMSE	
	z (mm)	$\dot{z}$ (mm/s)
UL_x	0.67	0.52
UL_y	1.20	0.81
LL_x	0.75	0.66
LL_y	1.04	0.98
LI_x	2.07	1.36
LI_y	1.45	1.13
TT_x	1.65	1.02
TT_y	2.53	1.64
TB_x	1.97	1.56
TB_y	1.90	1.58
TR_x	2.04	1.74
TR_y	2.03	1.55
Average	1.608	1.213

5.3. On-line tracking

The trained E-FNN estimates the MVs given the desired articulatory trajectories. It couples with the PID compensator for on-line adaptive control of the MSD system. The controller manipulates the MSD to infer the muscular activation patterns and to reproduce the desired articulatory trajectories in the 2-D articulatory synthesizer. Mermelstein's 2-D vocal tract model with a tract length of 17.5 cm is divided into 89 tube sections, with a uniform length of $\Delta x = 1.966 \times 10^{-3}$ m and a thickness of $\Delta y = 1 \times 10^{-2}$ m (Mermelstein, 1973; Boersma, 1998). The tube wall property is measured in the relax cheeks of a human adult speaker (Ishizaka et al., 1975; Birkholz and Jackèl, 2004), $M_0 = 21\,\text{kg/m}^2$, $B_0 = 8000\,\text{kg/m}^2\,\text{s}$, $K_0 = 845,000\,\text{kg/m}^2\,\text{s}^2$. The mass, damping, and stiffness parameters of the MSD manipulator in (1) are calculated as:

$$M = M_0 \Delta x \Delta y = 4.129 \times 10^{-4}\,\text{kg}, \tag{21}$$

$$B = B_0 \Delta x \Delta y = 0.157\,\text{kg/s}, \tag{22}$$

$$K = K_0 \Delta x \Delta y = 16.615\,\text{kg/s}^2. \tag{23}$$

The system state profile for the neutral or resting position are set as $\dot{z} = 0$, and $z = 0$. The gains of the PID compensator are set as $K_p = 25$, $K_i = 30$, and $K_d = 5$. The learning rate is $\eta = 0.005$.

6. Results and discussion

6.1. Articulatory trajectories

Smoothness is a main property of human speech articulation. We compare the reproduced articulatory trajectories in the proposed controller with the recorded EMA data of human speakers. Table 2 summarizes the RMSE of the controller during on-line tracking. The controller is able to manipulate the MSD and reproduce the desired position and velocity trajectories with high accuracy. Some TVs such as LI, TT, and TR demonstrate relatively higher error rates than others in Table 2. The observation suggests that the alveolar and the velar plosives possess a large amount of uncertainties in the CV sequences. In articulatory synthesis, the plosives are often independently generated using additional energy source at the constriction of the vocal tract during acoustic modeling (Birkholz et al., 2011). In practice, many researchers have suggested to reduce the degree of freedom or the error-prone TVs to increase the system efficiency (Birkholz, 2005; Ogata and Sonoda, 2003). For example, the jaw movement is considered as a secondary feature that smoothes the formant patterns during vowel production.

Treated as an inversion mapping model, the E-FNN is comparable to the trajectory mixture density network (TMDN) of Richmond (2009). Both models apply NNs to calculate the output probabilities based on the input vectors. In other

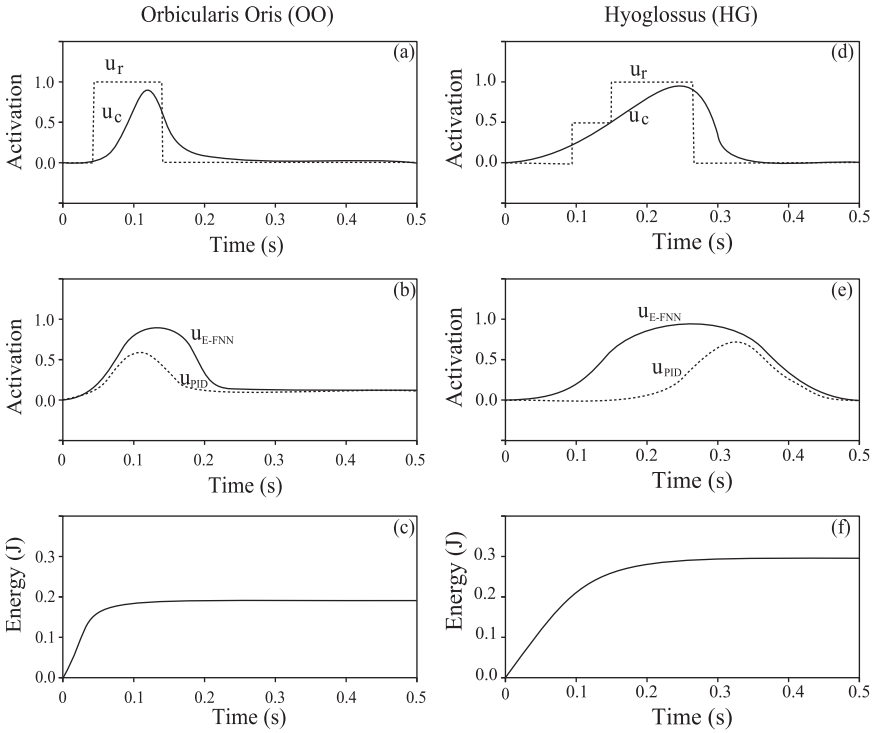

Fig. 5. The characteristics of motor activation and energy consumption in the *OO* and *HG* during [ɑ] reproduction. Upper panels (a) and (d) plot the reference control signal u_r (dashed line) and the inferred control signal u_c (solid line) in the fuzzy neural controller. Middle panels (b) and (e) plot the control signal of the E-FNN, $u_{\text{E-FNN}}$, and that of the PID compensator, u_{PID}, for the MSD system. Lower panels (c) and (f) plot the control effort or the energy consumption in the MSD system.

words, they use NNs to model the conditional probability density of the articulatory trajectories. However, there are a few differences. Firstly, the E-FNN maps between the motor variables and the articulatory trajectories, while the TMDN maps between the acoustic parameters and the articulatory parameters (i.e., priors, means and variances of the articulatory trajectories). They deal with different sets of feature vectors. Secondly, the E-FNN uses the recurrent neurons to account for the speech dynamics across the input feature vectors. Instead the TMDN uses a context window (10 frames), which is less effective than the recurrent layer. The effects of recurrent layer versus the context window have been investigated and reported by Schroeter and Sondhi (1994). Another major difference is the use of fuzzy logics to deal with the speech uncertainties and non-linearity in the proposed E-FNN model. Instead Richmond's TMDN model uses the fix-structured NNs. The later suffers when the training data is sparse, which is usually the case in speech data. Therefore, E-FNN attempts to revive the potentials of NNs using fuzzy logics in this speech study. In the literature, many methods have been proposed to model the non-linearity in the inversion mapping, and their accuracy are very similar (Toda et al., 2008). However, there is a lack of study on the underlying motor variables which causes the vocal tract shape changes and the variability in the surface acoustic-phonetic events. In future studies, we are interested to apply the E-FNN structure for the inversion mapping experiments in comparison with other existing methods.

6.2. Muscular activations

For adaptive control, it is beneficial to extract the underlying articulatory commands, the MVs, to explain the dynamics of the TVs. Fig. 5 shows the motor control signals of two MVs, *OO* and *HG*, in the proposed controller for the reproduction of [ɑ] sequence. Fig. 5(a) and (d) plots the reference control signal u_r (dashed line) and the inferred control signal u_c (solid line), where $u_c = u_{\text{E-FNN}} - u_{PID}$ (Section 4, (14)). Fig. 5(b) and (e) plots the control input of the E-FNN, $u_{\text{E-FNN}}$ and that of the PID compensator, u_{PID}. The proposed E-FNN controller demonstrates better performance than the linear compensator in terms of inversion accuracy. The motor activation data agrees with the

measured EMG data of Baer et al., where the phones have distinctive targets in the control space (Baer et al., 1988). Four things are observed in the data.

1. The motor activation resembles the step response in the MSD. Ogata and Sonoda (2003) have previously used time-invariant linear systems excited by impulse trains to reproduce the velocity profiles of the speech articulators, which resembles the idea of motor inversion in this study.
2. The delay between the MV onset and the TV onset (at the boundary of the gate-like reference control signal) is roughly 30–70 ms, which corresponds to the reaction time from muscle activation to articulatory motion in human speech production (Birkholz et al., 2011).
3. The controller is able to model the non-linearities of the articulators during CV production, which is evidenced by the smooth *HG* activation curve from [b] to [ɑ].
4. In the articulatory synthesis model, the controller moves the articulator back to the neutral position without oscillation, which mimics the human speech articulation (Saltzman and Munhall, 1989; Perrier and Ostry, 1996).

6.3. Speech motor control for articulatory synthesis

The control model can be integrated with the anatomical and the acoustic models in a full articulatory synthesizer for TTS applications. However, the control model needs to specify the prosody information besides the phone sequences, such as the speaking rate in the input text. One difficulty is that the mapping from the articulatory trajectories to the acoustic sound is not strictly "one-to-one", where there can be more than one vocal tract-cord configuration that produce the same acoustic sound in the synthesis system. The issue can be simplified by balancing the trade-off between the articulatory effort and the acoustic distinctiveness using an optimal control strategy (Kröger et al., 2009; Perrier et al., 2005). For speech motor control, the EMG measures are the combined results of efferent and afferent influences from the biomechanical properties of the muscular structures (Buchaillard et al., 2009; Perrier et al., 2005). The muscular forces can alter the position and velocity of the articulators. In the proposed controller, we are able to examine the excitation pattern and calculate the input energy of the dynamic MSD system,

$$E = \int_{t_0}^{t_s} u(t)\dot{z}(t)dt, \tag{24}$$

where $u(t)$ is the input force, u_c, $\dot{z}(t)$ is the velocity of the tract wall at time t, and t_s is the sampling time. Fig. 5(c) and (f) plots the energy consumption in joules (J) of two MVs, *OO* and *HG*, during [ɑ] production. Energy rises abruptly for *OO* in the lips when producing the labial plosive [b], but the overall measure is lower compared to *HG* when producing the low vowel [ɑ].

In the proposed controller, it is possible to calculate the overall control effort of the MVs in the articulatory synthesizer, where a "minimum energy" criterion can be embedded for optimal control (Kawato et al., 1990). The criterion is analogous to the speaker-oriented "minimum articulatory cost" in the functional phonology of speech production, where the speaker seeks to minimize the articulatory effort while maintaining the distinctiveness of the acoustic sounds during speech production (Browman and Goldstein, 1992; Boersma, 1998). However, the control energy in Fig. 5(c) and (f) are only relative measures since the MVs are normalized to [0, 1]. For example, the activation of $GGp(=1)$ exerts a force of 25.82 N during [i] production, while the activation of $SG(=1)$ exerts a force of 6.9 N (Buchaillard et al., 2009). If used for optimal speech motor control, the MVs should have different prominence. However, Perrier et al. (2003) argued that the optimized control is not necessary for the smoothly varying articulatory movements. They also showed that the bio-mechanical characteristics of the speech articulators alone can answer for such kinematic property. In this study, the MVs are extracted for the 2-D articulator with MSD based tube wall. Therefore, it has limited capability when evaluating the optimal control strategies. Nonetheless, the proposed controller is the first step toward an automatically controlled articulatory synthesizer. The inferred MVs can also provide an alternative set of motor features to describe the acoustic-phonetic events for improved speech recognition. For example, Mitra et al. used the articulatory synthesizer to prepare a codebook of acoustic-to-articulatory data pairs, and they showed that the inferred articulatory features increased the robustness toward noise contamination and speaker variations in the speech recognition systems (Mitra et al., 2011). We are writing another paper using the articulatory based features in speech recognition.

7. Conclusion

The shape, position, and movement of the articulators are the immediate targets of human speech production (Saltzman and Munhall, 1989; Kelso et al., 1986). Reproducing these smooth and natural trajectories is critical for high quality articulatory speech synthesis. This paper presents an adaptive fuzzy neural controller, which tracks the measured articulatory trajectories in the form of TVs and infers the underlying muscular excitation patterns in the form of MVs. Major characteristics of the proposed adaptive fuzzy neural controller are as follows.

1. The E-FNN controller models the inverse dynamics between the motor commands and the tract variables in an off-line mode, where the structure and parameters of the neural topology are automatically and dynamically determined on the speech data.
2. The controller deals the uncertainties and the non-linearities in the MSD system using an adaptive control law.
3. Compared to the fixed structured NNs, the self-adaptation and learning ability of the E-FNN controller is more adequate to model the dynamics of the articulators.

The proposed controller demonstrates good tracking performance on the CV sequences. It reproduces the smooth and bell-shaped articulatory trajectories, and it retrieves the motor activations patterns in the vocal tract. The mapping characteristics between the MVs and the TVs are useful for speech motor control in articulatory synthesis, they are also useful for applications in automatic speech recognition.

References

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

Baer, T., Alfronso, P.J., Honda, K., 1988. Electromyography of the tongue muscle during vowels in /epvp/ environment. Annual Bulletin of the Research Institute of Logopaedics and Phoniatrics 22, 7–19.

Bailly, G., 1997. Learning to speak. Sensori-motor control of speech movements. Speech Communication 22, 251–267.

Birkholz, P., 2005. 3-D Articulatory Speech Synthesis. Ph.D. thesis. University of Rostock, Rostock, Germany.

Birkholz, P., Jackèl, D., 2004. Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In: Interspeech, ISCA.

Birkholz, P., Jackel, D., Kroger, B., 2007. Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1218–1226.

Birkholz, P., Kroger, B., Neuschaefer Rube, C., 2011. Model-based reproduction of articulatory trajectories for consonant–vowel sequences. IEEE Transactions on Audio, Speech, and Language Processing 19, 1422–1433.

Boersma, P., 1998. Functional phonology: formalizing the interactions between articulatory and perceptual drives. Ph.D. thesis. University of Amsterdam.

Browman, C.P., Goldstein, L., 1992. Articulatory phonology: an overview. Phonetica 49, 155–180.

Buchaillard, S., Perrier, P., Payan, Y., 2009. A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning. Journal of the Acoustical Society of America 126, 2033–2051.

Cook, P., 1990. Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing. Master's thesis. Stanford University. Stanford, California.

van den Doel, K., Ascher, U., 2008. Real-time numerical solution of Webster's equation on a nonuniform grid. IEEE Transactions on Audio, Speech, and Language Processing 16, 1163–1172.

Duck, F.A., 1990. Physical Properties of Tissues: A Comprehensive Reference Book. Academic Press, London.

Er, M.J., Gao, Y., 2003. Robust adaptive control of robot manipulators using generalized fuzzy neural networks. IEEE Transactions on Industrial Electronics 50, 620–628.

Fang, Q., 2009. A Study on Construction and Control of a Three-dimensional Physiological Articulatory Model for Speech Production. Ph.D. thesis. School of Information Science, Japan Advanced Institute of Science and Technology.

Feldman, A.G., 1986. Once more on the equilibrium-point hypothesis (lambda model) for motor control. Journal of Motor Behavior 18, 17–54.

Flanagan, J., Ishizaka, K., Shipley, K., 1975. Synthesis of speech from a dynamic model of the vocal cords and vocal tract. The BELL System Technical Journal 54, 485–506.

Gao, Y., Er, M.J., 2003. Online adaptive fuzzy neural identification and control of a class of MIMO nonlinear systems. IEEE Transactions on Fuzzy Systems 11, 462–477.

Gao, Y., Er, M.J., 2005. An intelligent adaptive control scheme for postsurgical blood pressure regulation. IEEE Transactions on Neural Networks 16, 475–483.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., 1993. Inverse dynamics of speech motor control. In: NIPS, pp. 1043–1050.

Ishizaka, K., French, J.C., Flanagan, J.L., 1975. Direct determination of vocal tract wall impedance. IEEE Transactions on Acoustics, Speech and Signal Processing 23, 370–373.

Jordan, M.I., 1986. Serial order in behavior: a parallel distributed processing approach. Technical Report 8604. University of California, Institute for Cognitive Science, San Diego.

Kawato, M., Maeda, Y., Uno, Y.R.S., 1990. Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. Biological Cybernetics 62, 275–288.

Kelso, J.A.S., Saltzman, E.L., Tuller, B., 1986. The dynamical perspective on speech production: data and theory. Journal of Phonetics 14, 29–59.

Kim, K., Gomi, H., 2007. Model-based investigation of control and dynamics in human articulatory motion. Journal of System Design and Dynamics 1, 558–569.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition. Journal of the Acoustical Society of America 121, 723–742.

Kröger, B., Graf-Borttscheller, V., Lowit, A., 2008. Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In: Interspeech, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia.

Kröger, B., Kannampuzha, J., Neuschaefer-Rube, C., 2009. Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809.

Kröger, B., Schroder, G., Opgen-Rhein, C., 1995. A gesture-based dynamic model describing articulatory movement data. Journal of the Acoustical Society of America 98, 1878–1889.

Lin, C.M., Li, H.Y., 2012. TSK fuzzy CMAC-based robust adaptive backstepping control for uncertain nonlinear systems. IEEE Transactions on Fuzzy Systems 20, 1147–1154.

Löfqvist, A., Gracco, V., 2002. Control of oral closure in lingual stop consonant production. Journal of the Acoustical Society of America 111, 2811–2827.

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

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., Goldstein, L., 2011. Articulatory information for noise robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 19, 1913–1924.

Nelson, W.L., 1983. Physical principles for economies of skilled movements. Biological Cybernetics 46, 135–147.

Ogata, K., Sonoda, Y., 2003. Reproduction of articulatory behavior based on the parameterization of articulatory movements. Acoustic Science and Technology 24, 403–405.

Perrier, P., Ma, L., Payan, Y., 2005. Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue. In: 9th European Conference on Speech Communication and Technology, pp. 1041–1044.

Perrier, P., Ostry, D.J., 1996. The equilibrium point hypothesis and its application to speech motor control. Journal of Speech and Hearing Research 39, 365–378.

Perrier, P., Payan, Y., Zandipour, M., Perkell, J., 2003. 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, 1582–1599.

Richmond, K., 2009. Preliminary inversion mapping results with a new EMA corpus. In: Interspeech'09.

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

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

Slotine, J.J.E., Li, W., 1991. Applied Nonlinear Control. Prentice-Hall, Upper Saddle River, NJ.

Toda, T., Black, A., Tokuda, K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Communication 50, 215–227.

Wang, L.X., 1997. A Course in Fuzzy Systems and Control. Prentice-Hall, NJ.

Wrench, A., 1999. The MOCHA-TIMIT articulatory database.

Wu, S., Er, M.J., Gao, Y., 2001. A fast approach for automatic generation of fuzzy rules by generalized dynamic fuzzy neural networks. IEEE Transactions on Fuzzy Systems 9, 578–594.