USING STATE FEEDBACK TO CONTROL AN ARTICULATORY SYNTHESIZER

Ian S. Howard[1] & Peter Birkholz[2]

[1]Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, PL4 8AA, UK.
UK Email: ian.howard@plymouth.ac.uk
[2]Institute of Acoustics and Speech Communication, TU Dresden, 01062 Dresden, Germany.
Email: peter.birkholz@tu-dresden.de

Abstract: Here we consider the application of state feedback control to stabilize an articulatory speech synthesizer during the generation of speech utterances. We first describe the architecture of such an approach from a signal flow perspective. We explain that an internal model is needed for effective operation, which can be acquired during a babbling phase. The required inverse mapping between the synthesizer's control parameters and their auditory consequences can be learned using a neural network. Such an inverse model provides a means to map output that occur in an acoustic speech domain back to an articulatory domain, where it can assist in compensatory adjustments. We show that it is possible to build such an inverse model for the Birkholz articulatory synthesizer for vowel production. Finally, we illustrate the operation of the inverse model with some simple vowels sequences and static vowel qualities.

1 Introduction

In order to speak, we need to move the speech articulators in an appropriate fashion. Therefore, at its lowest mechanical level, speech production can be considered to be a motor task that leads to acoustic consequences. Of course, it is the latter which is of primary interest to a listener. It is well established that if during speech production, articulator position is perturbed, human speakers generate compensatory movements to counteract the disturbance, such as those see when mechanical perturbations are made to the jaw [1]. Similarly, changes to auditory feedback that affect vowel quality can also be compensated [2]. Such observed compensatory behavior suggests that some kind of feedback control mechanisms operate in the human speech production process that make use of both proprioceptive and auditory feedback.

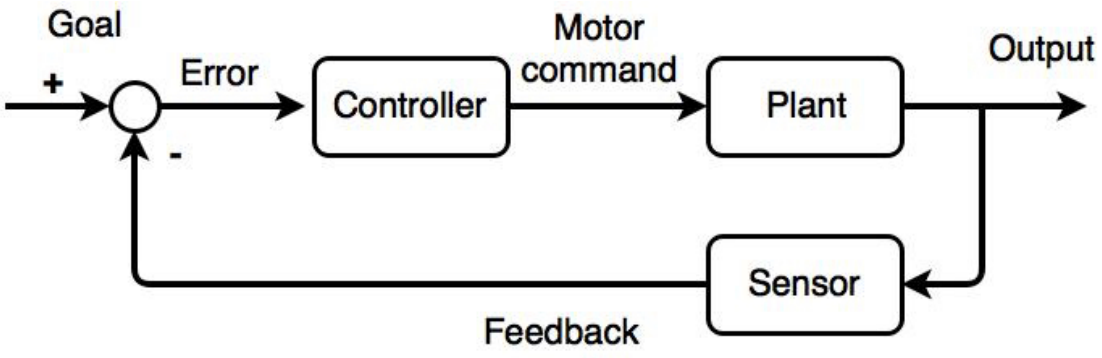

Fig. 1 Using output feedback control, the sensory consequences scaled by a gain factor are used to modify the control input by comparing it with the goal to calculate error and this is then used in an attempt to get the plant to meet the required goals. This scheme also has the ability to compensate for disturbances.

2 Feedback control

Controlling any real physical system, including the human speech apparatus, involves not only dealing with the dynamics of the moving parts, but also dealing with any unpredictable disturbances that may occur. The field of control engineering provides a useful means to understand such issues, and also offers computational solutions to these kinds of problems. Feedback control (Fig. 1) is often used in engineering systems to stabilize operating goals when noise is present. For such a paradigm to operate effectively, the feedback gain needs to be set sufficiently high to achieve good performance – such as fast movement to targets and good compensation to disturbance, but it also needs to be chosen to avoid the resulting system from becoming unstable.

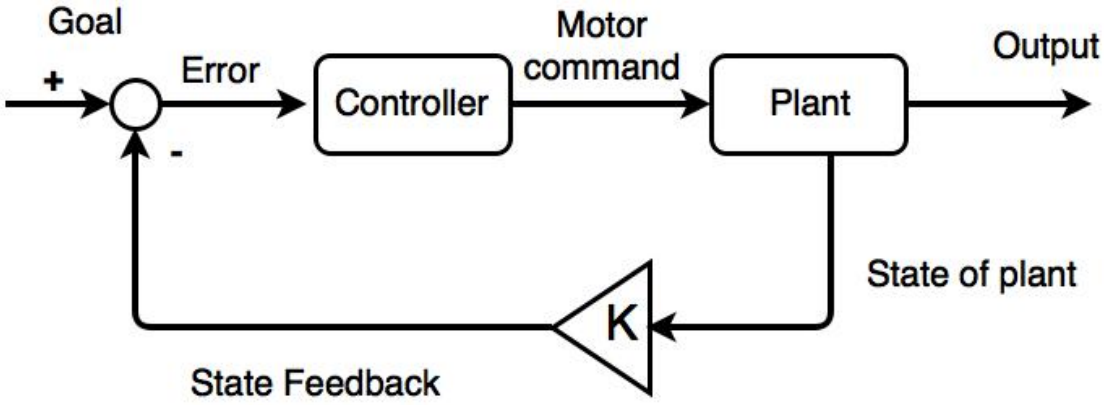

Fig. 2 Using direct state feedback control. Lower path shows the state feedback signal flow, which includes multiplication by the feedback gain vector K. In practice, an observer (also known as a forward model) is often used to estimate the system state.

Control can often be improved by making use of full state feedback, and not just the output of the system., as shown in Fig. 2. Such a state feedback control (SFC) architecture uses the full estimated state of the system, which is generally a vector and not just a single scalar value. This state is then weighted appropriately and used to generate a scalar control signal corresponding to the error between the desired and estimated states. This error is then used to correct the plant to enable it to follow the desired goals. In practice, a state estimation mechanism may be needed, which can be realized using an observer, since not all of the systems states may be directly available. Such an observer also provides an elegant way to deal with the issue of delayed sensory feedback.

State feedback control has been recently proposed as a good framework to understand observed phenomena in human speech production [3]. Following on from this work, state feedback has also been used to control phonation pitch in a simplified model of the vocal folds [4]. In this work, the larynx is modelled as a single damped mass-spring system and it generates auditory and somatosensory output. The auditory and somatosensory systems received state input from a state estimator that are used to calculate errors in their respective modalities and then are mapped back for use in the control domain. These signals are then used to update estimates of laryngeal state. This is illustrated in Fig. 3. The authors showed that their model was able to compensate for perturbations made to auditory feedback.

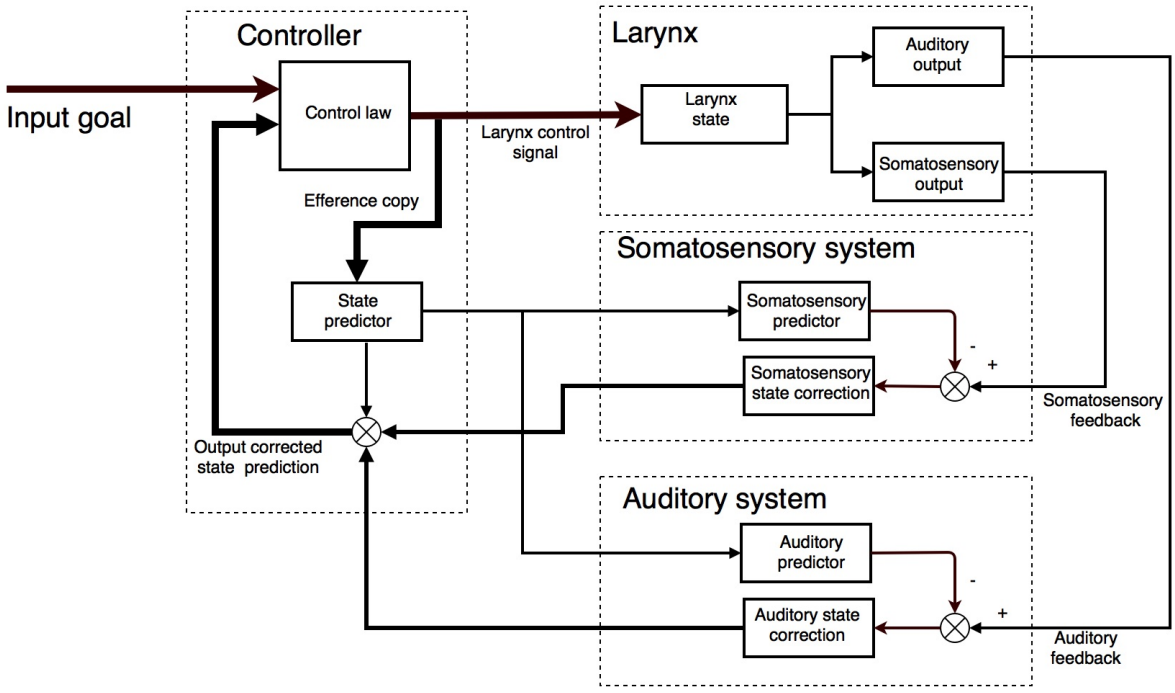

Fig. 3 SFC architecture used for vocal larynx control. Diagram redrawn from the work by Houde et al. [4]. This scheme makes use of forward models to predict both somatosensory and auditory consequences using the control input to the larynx. In addition, it uses inverse models to map somatosensory and auditory errors back to motor representation.

Here we consider a how to implement a state feedback control scheme to operate the Birkholz articulatory speech synthesizer [5]. We propose to directly drive the vocal tract articulators with position trajectories (as is often done in software articulatory speech synthesizers), and therefore do not need to address the issue of the control that arise from articulator dynamics, or make use of an observer to predict system state (although such features could be easily incorporated into the paradigm). However, this assumption lets us make use of the specified articulator positions as an estimate of the vocal tract's proprioceptive state. Nevertheless, we still need to make use of an indirect estimate of articulatory state made on the basis of acoustic output. Such an estimate can be made by employing an inverse model that maps acoustic sensory consequences back to the corresponding articulatory configuration. Therefore, within this feedback scheme, both proprioceptive and acoustic elements in the state vector contribute to the correction process when speech production is disturbed.

In these preliminary experiments, we investigate how to develop inverse models that map between auditory control parameter domains for vowel production. Although the auditory inverse model using by Houde [4] is used to map back auditory error (Fig 3), here we consider using an inverse model to map the auditory output of the synthesizer to the corresponding articulatory control parameters and then generate the corresponding error in the articulatory domain, as illustrated by the architecture is illustrated in Fig. 4. In this arrangement, if articulator position is perturbed, both proprioceptive and acoustic error will contribute to the correction of the articulatory system.

3 Methods

Training an inverse model which maps acoustic consequences back to articulatory control signals is easy to achieve. To design, implement and train the inverse model, we follow a similar approach as one used previously [6], [7]. In short, all that is necessary is to drive the vocal tract synthesiser by appropriate pseudorandom input, such as parameter trajectories corresponding to speech babble. This subsequently leads to the generation of corresponding speech output. In this scenario, both the articulatory control signals and their acoustic consequences are available and can be used in a supervised learning scheme to train a neural network that maps between acoustic consequences and the articulatory control signals responsible for them. This is shown in Fig. 5.

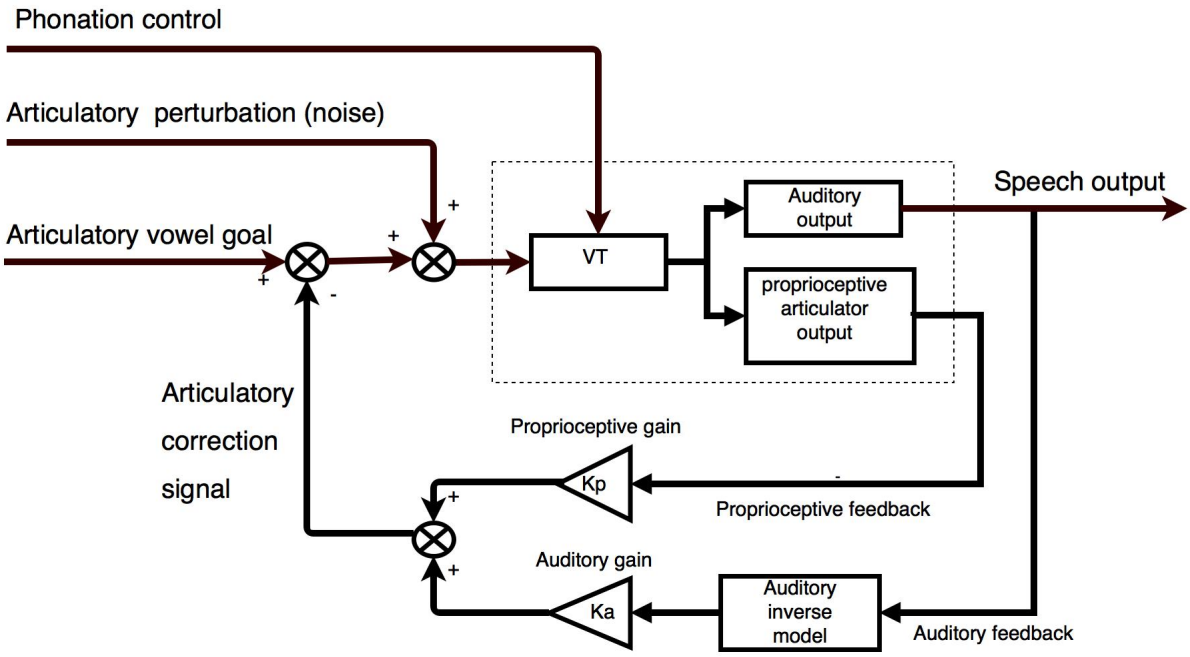

Fig. 4 signal flow diagram for direct kinematic control of vocal tract articulators. Here articulator state is obtained directly from the kinematic input. However, articulatory state estimated on the basis of acoustic output needs to employ an inverse model to map between the auditory and articulator domains.

To train an inverse model, a babble generator was run to generate repeating sequences of 16 vowels for a male speaker. Cosine interpolation was used between locations vowel qualities resulting in a 14-parameter articulatory control vector specified every 5ms. In addition, the glottal parameters were appropriately specified and fundamental frequency for each vowel region was set at random between 110 and 130 Hz. In total about 75 seconds of articulator trajectory data was generated. These parameter trajectories were used to generate output speech which was subsequently analyzed acoustically. The analysis was based on an auditory filter bank [8]. After suitable down sampling, this resulted in a 16-channel frequency frame data vector every 5ms. The resulting vocal tract parameter trajectories and their corresponding down sampled filter bank output are shown in Fig. 6.

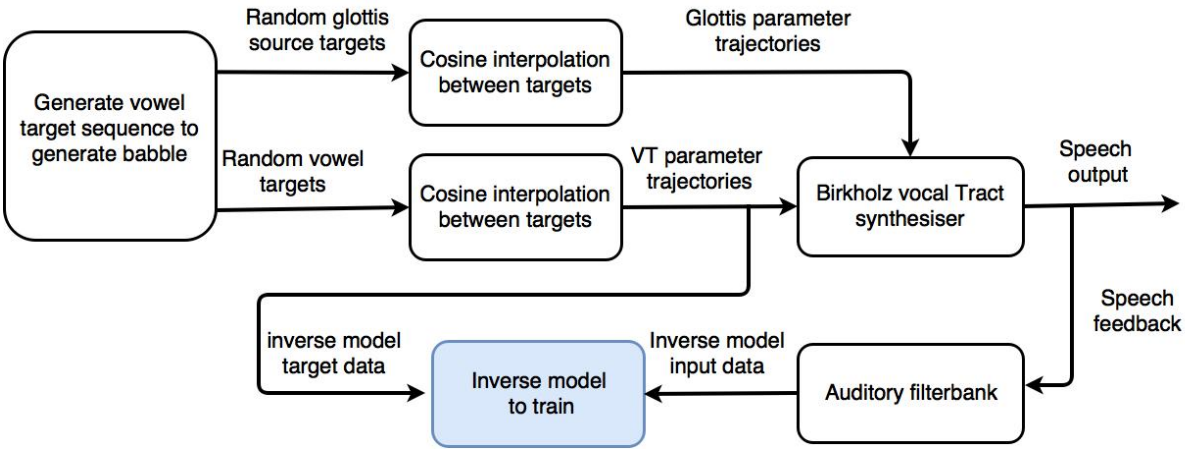

Fig. 5 Training the inverse model. It is possible to generate the input and output data need to estimate an inverse model by running the vocal apparatus to produce speech babble. This is achieved by generating random vocal parameter trajectories using a babble generator, and this signal becomes the output training target for the inverse model. It is also used to drive the vocal tract synthesizer and the corresponding acoustic output is then fed into an auditory filter back. This generates an acoustic representation of the sensory consequences of the motor action that becomes the input training data for the inverse model.

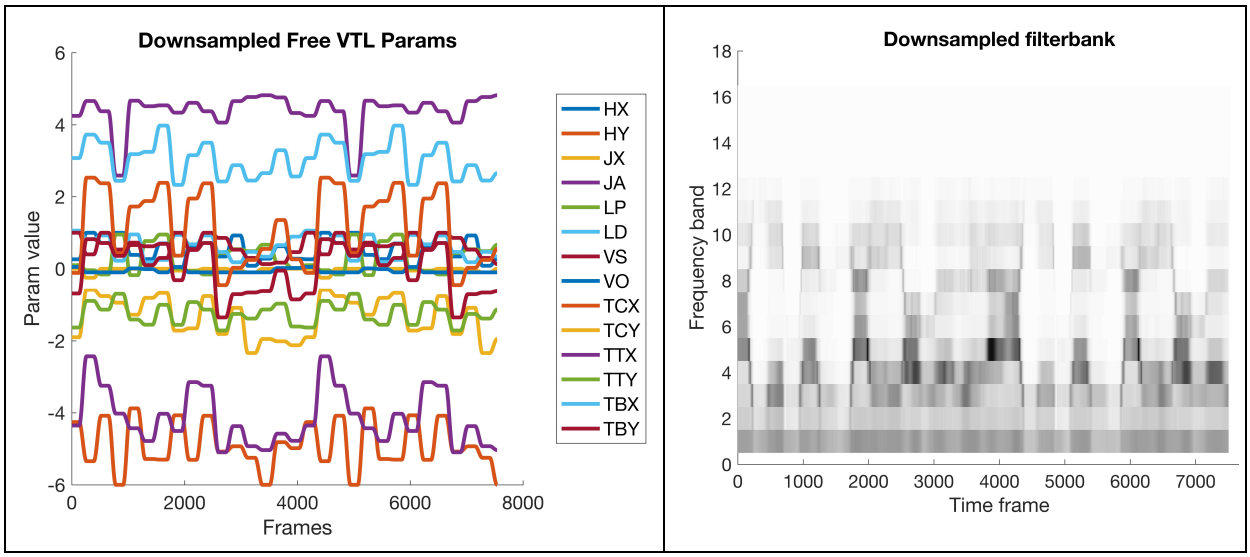

Fig. 6 Inverse model training data. Left panel shows target control parameter trajectories made by cosine interpolation between vowel targets, resulting in babble consisting of 16 vowel qualities. Right panel shows corresponding output from the auditory filter bank.

To realize an inverse model, a Matlab implementation of a multi-layer perceptron (MLP) was used [9]. The input to the inverse model consisted of 10 centered adjacent filter bank frames spanning 50ms in time in total, and the MLP had 40 hidden units and 14 linear outputs. Input and output data patterns were normalized by subtracting their mean value and dividing by their standard deviation. The MLP was trained using back-propagation with conjugate gradient descent. Training the inverse model involved 2000 passes over the data set. After training, the inverse model was used in recognition mode, its output was un-normalized by multiplying by the training set standard deviation and adding the training set mean value.

4 Results

The inverse model was tested by observing the predicted parameter control trajectories and also by re-synthesizing input speech. This was achieved by passing speech utterances generated by the synthesizer through the acoustic analysis inverse model and finally to the synthesizer. Evaluations were carried by observation of the corresponding filter bank outputs and listening tests. Subjective inverse model performance was good, and the resynthesized speech was almost indistinguishable from the original synthesized input speech.

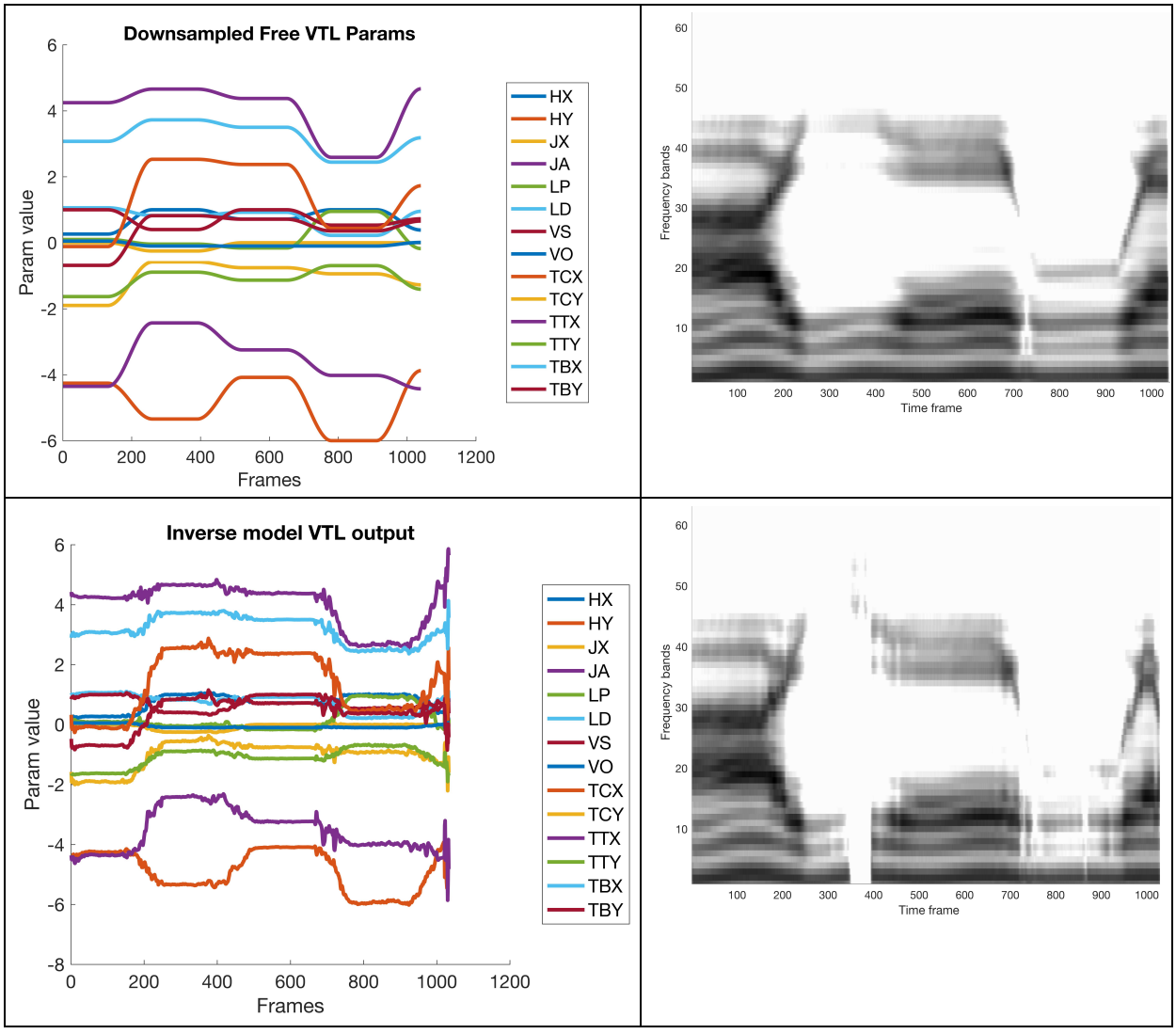

Fig. 7 Example sequence of 5 vowels to illustrate the operation of the inverse model. Upper left shows vocal tract parameter trajectories and upper right shows corresponding filter bank spectrogram of resulting synthesized speech output. Channels represent a frequency range of 0-3kHz. Lower left shows vocal tract parameter trajectories estimated by the inverse model and lower right shows corresponding filter bank spectrogram of resulting from re-synthesizing speech.

A good correspondence in input and can been seen by comparing the respective speech spectrograms shown in Fig. 7. We note that the small deviations in the parameter trajectories arise because that fundamental frequency contour in the testing data was random and differed from that experienced during training. Comparisons of target and inverse mode reconstructed

parameter trajectories for static vowels are shown in Fig. 8. Again, the glitches in the trajectories arise due to fundamental frequency effects.

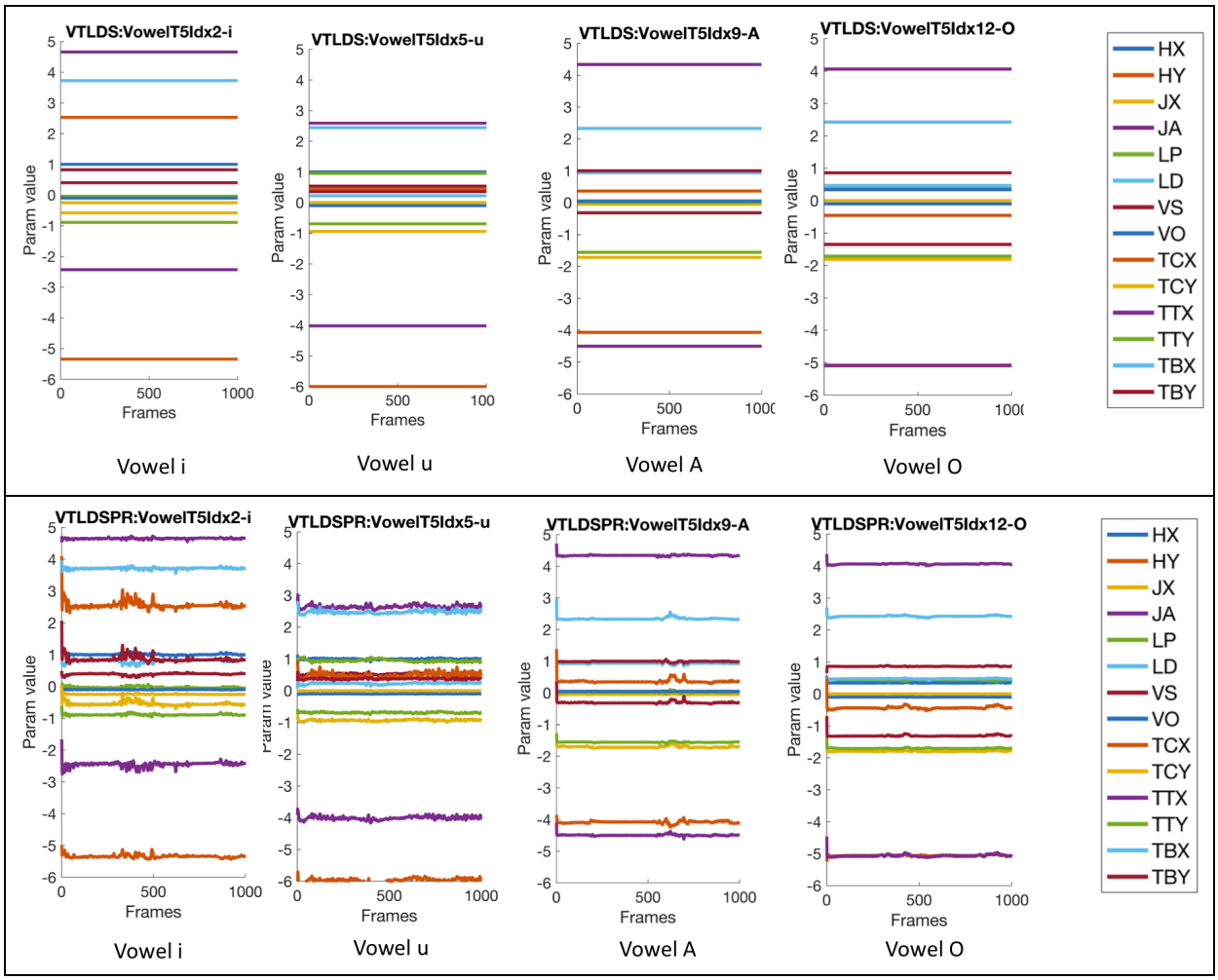

Fig. 8 Upper panel shows target static vowel vocal tract parameter trajectories. Lower panel shows the corresponding inverse model output vocal tract parameters for the same 4 target vowels.

5 Discussion

In this paper, we consider operating the Birkholz articulatory speech synthesizer using state feedback control and as a first step in this process, investigated training inverse models that can map between auditory and control parameter domains. To do so, we drive the articulatory synthesizer directly from target trajectories that specify articulate a location. such trajectory is completely specified synthesizer behavior. Here we have avoided many important issues. For example, we have not addressed the issue of state of estimation at any great length. Neither have we considered the issue of temporal delay, although both of these issues are clearly important.

In more sophisticated future simulation of the vocal apparatus, force control could be used and the dynamics of the articulator taken into account. In such a case, it would be necessary to model control of the dynamical system, rather than making use of a direct kinematic control as adopted here. Going one stage further, approaches such as the task dynamic model also attempt to model task directed behaviors of the vocal apparatus, such as the importance of

area functions in the vocal tract. To incorporate state feedback control in such approaches, it is also necessary to take into account the transformations between task and articulator dynamics. and indeed, work in this area has already been carried out by Ramanarayanan and colleagues [10].

Finally, although state space feedback control is promising way to explain and understanding human speech production [3], we note that in the field of sensori-motor control, the related field of optimal control [11] currently represents the best theoretical framework to amount for observation of human movement behavior, and will no double have much to offer the field of speech production too.

6 References

[1] S. TREMBLAY AND D. SHILLER, "Somatosensory basis of speech production," *Nature*, 2003.

[2] J. F. HOUDE, "Sensorimotor Adaptation in Speech Production," *Science*, vol. 279, no. 5354, pp. 1213–1216, Feb. 1998.

[3] J. F. HOUDE AND S. S. NAGARAJAN, "Speech production as state feedback control.," *Front Hum Neurosci*, vol. 5, p. 82, 2011.

[4] J. F. HOUDE, C. NIZIOLEK, N. KORT, Z. AGNEW AND S. S. NAGARAJAN, "Simulating a state feedback model of speaking", *Seminar on Speech*, Cologne, 2014.

[5] P. BIRKHOLZ, D. JACKEL, AND B. J. KRÖGER, "Construction and Control of a Three-Dimensional Vocal Tract Model," presented at the 2006 IEEE International Conference on Acoustics Speed and Signal Processing, 2006, vol. 1.

[6] I. HOWARD AND M. HUCKVALE, "Training a vocal tract synthesizer to imitate speech using distal supervised learning," *Proc SPECOM*, 2005.

[7] I. HOWARD AND M. HUCKVALE, "'Learning to Control an Articulator Synthesizer by Imitating Real Speech," *ZASPIL*, 2004.

[8] M. SLANEY, *An Efficient Implementation of the Patterson–Holdsworth Auditory Filter BanK*. Perception Group, Tech. Rep, 1993, 1993.

[9] I. T. Nabney, *Nabney: Netlab: Algorithms for Pattern Recognition. 2004 - Google Scholar*. London.

[10] V. RAMANARAYANAN, B. PARRELL, L. GOLDSTEIN, S. NAGARAJAN, AND J. HOUDE, "A New Model of Speech Motor Control Based on Task Dynamics and State Feedback," presented at the Interspeech 2016, 2016, vol. 2016, pp. 3564–3568.

[11] E. TODOROV AND M. I. JORDAN, "Optimal feedback control as a theory of motor coordination.," *Nat Neurosci*, vol. 5, no. 11, pp. 1226–1235, Nov. 2002.