Spatial-to-joint coordinate mapping in a neural model of speech production

Bernd J. Kröger[1], Peter Birkholz[2], Jim Kannampuzha[1], Christiane Neuschaefer-Rube[1]

[1] *Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen, 52074 Aachen, Germany, Email: bkroeger@ukaachen.de*

[2] *Department of Computer Science, University of Rostock, 18051 Rostock, Germany, Email: piet@informatik.uni-rostock.de*

Abstract

The mapping from a high-level to a low-level motor representation, i.e. from spatial-to-joint motor coordinates is modeled on the basis of a one-layer feed-forward neural network and supervised learning using articulatory and acoustic data generated by a three dimensional articulatory speech synthesizer.

Introduction

Spatial coordinate representations constitute a central level in sensorimotor control of movements [1]. In the case of speech motor control spatial coordinates are also labeled as tract variables [2]. The tract variable system directly describes degree and location of vocal tract constrictions and thus allows a high-level goal-oriented description of speech gestures as needed for speech motor planning. Since motor execution and low-level motor control needs a joint coordinate description of movements – e.g. movement of tongue and lower lip relative to jaw – a transformation from high-level spatial coordinates to low-level joint coordinates is central in speech motor control. This transformation can be implemented numerically using the task-dynamics approach [2] or in a much simpler way using neural nets [3, 4]. In this paper a neural network approach for the spatial-to-joint coordinate mapping using a high quality three dimensional articulatory speech synthesizer [5] is described.

The 3D articulatory speech synthesizer

The 3D articulatory speech synthesizer or articulatory-acoustic model [5] is controlled by a set of 10 articulatory or joint coordinate motor parameters (Tab. 1) defining the position of each model articulator, e.g. lips, tongue, jaw, and larynx. The spatial coordinate parameters or tract variables (Tab. 2) are based on flesh point locations calculated from the 3D model in the cranial x-y-coordinate system (Fig. 1).

The spatial-to-joint coordinate mapping

Training the spatial-to-joint coordinate net is successfull using a one-layer feedforward network (Fig. 2) and a training set comprising all combinations of minimum and maximum position values (i.e. 0 and 1) of the normalized 10 articulatory parameters. According to physiolgical constraints for the interdependencies of tongue body and tongue tip articulation, 5 subsets of relative min-max position values are used in the case of the parameters TBA, TBL, TTA, TTL leading to a maximum distance of 0.5 (Fig. 3). Furthermore all parameter combinations leading to physiologically impractical simultaneous double closures of tongue tip and tongue body with the palate are eliminated resulting in a training set comprising 4608 training patterns (i.e. articulatory states). An extension of this set of articulatory states by combining more position values for each articulator mainly blows up the number of articulatory states without leading to a significant enhancement of the quality of the net.

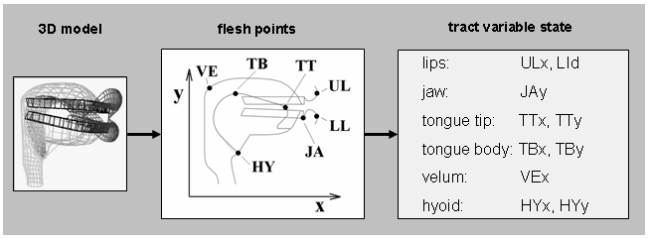

Fig. 1: Generation of tract variable values using the 3D model.

Tab. 1: List of articulatory parameters (joint coordinate motor parameters)

ABBR.	NAME OF ARTICULATORY PARAMETER
JAA	lower jaw angle
TBA	tongue body angle
TBL	tongue body horizontal location
TTA	tongue tip angle
TTL	tongue tip horizontal location
LIH	relative lip height
LIP	lip protrusion
VEH	velum height
HLH	hyoid horizontal location
HLV	hyoid vertical location

Relative lip height means: lip height relative to jaw.

Tab. 2: List of tract variables (spatial coordinate motor parameters)

ABBR.	NAME OF TRACT VARIABLE
ULx	upper lip horizontal position
JAy	lower jaw vertical position
TTx	tongue tip horizontal position
TTy	tongue tip vertical position
TBx	tongue body horizontal position
TBy	tongue body vertical position
VEx	velum horizontal position
HYx	hyoid horizontal position
HYy	hyoid vertical position
LId	lips vertical distance

100000 cycles batch training are sufficient for obtaining an error smaller than 10% for predicting joint coordinate patterns. Using the trained spatial-to-joint coordinate net a variation of each single tract variable – while keeping the values of all other tract variables constant – leads to a direct control of position and degree of vocal tract constrictions.

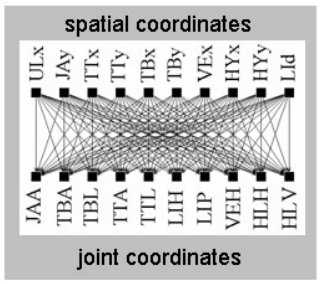

Fig. 2: Spatial-to-joint coordinate network.

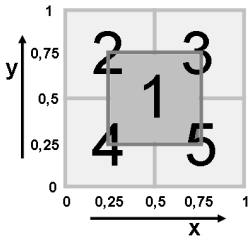

Fig. 3: The 5 rectangles represent 5 subsets of position value combinations for 4 physiologically constrained spatial parameters (TBA, TBL, TTA, TTL) for establishing a training set for the spatial-to-joint coordinate network. These 5 subsets are combined with the min-max combinations of all other spatial parameters. The distance of physiologically constrained position values does not exceed 0.5 within the training set, while the distance is 1 in the case of all other position value combinations for JAA, LIH, LIP, VEH, HLH, HLV.

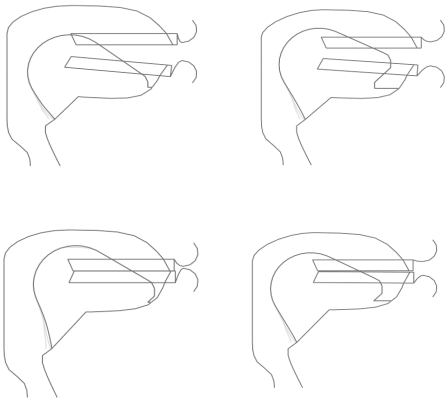

Fig. 4: Left side: Control of articulation by joint coordinate parameters: Exclusive variation of the parameter JAA keeping all other joint coordinate parameters constant. Right side: Control of articulation by spatial coordinate parameters: Exclusive variation of the parameter JAy keeping all other spatial coordinate parameters constant. (light-gray lines: lateral tongue contours)

An example for control of articulation by spatial parameters is given in Fig. 4 (right side): Lower jaw position is allowed to vary while all other spatial are fixed at a mid position value within their total range. Lip, tongue body, and tongue tip position remains relatively constant within the cranial x-y-coordinate system. In the case of control of articulation by joint coordinates (Fig. 4, left side) the articulators lower lip, tongue body, and tongue tip adopt the jaw movement.

Since location and degree of vocal tract constrictions are maintained using spatial coordinates for control of articulation, the model is capable of realising motor equivalence or compensatory articulation. An example for modeling motor equivalence in the case of labial, apical, or dorsal vocal tract closure using different lower jaw positions is given in Fig. 5: Despite different positoning of lower jaw the vocal tract closure is maintained in all three cases.

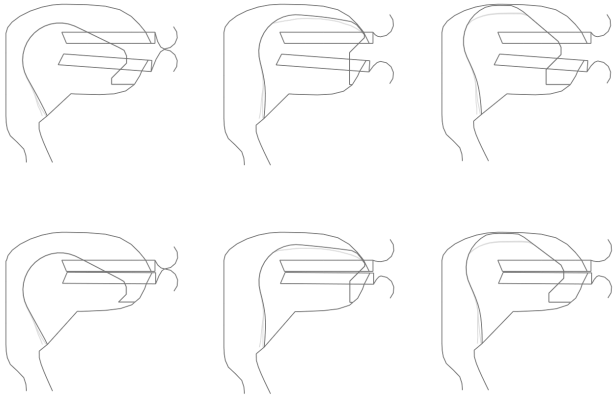

Fig. 5: Production of labial, apical, and dorsal vocal tract closure using different lower jaw positions (low and high). The same set of spatial coordinate parameters (i.e. tract variables) is used for each type of vocal tract closure with exception of the tract variable JAy. (light-gray lines: lateral tongue contours)

Conclusion and further work

Modeling of the spatial-to-joint coordinate mapping is successful using our three dimensional articulatory speech synthesizer. This work serves as a basis for establishing a complete model of sensorimotor control of speech production.

Acknowledgment

This work was supported in part by the German Research Council DFG under Grant KR1439/10-1 and under Grant JA1476/1-1.

Literature

[1] Kandel ER, Schwartz JH, Jessell TM (2000) *Principles of neural science.* MacGraw-Hill, New York

[2] Saltzman EL, Munhall KG (1989) A dynamic approach to gestural patterning in speech production. *Ecological Psychology* 1, 333-382

[3] Guenther FH (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review* 102, 594-621

[4] Guenther FH, Ghosh SS, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96, 280-301

[5] Birkholz P, Jackel D, Kröger BJ (2006) Development and control of a 3D vocal tract model. *Proceedings of the IEEE International conference on Acoustics, Speech, and Signal Processing ICASSP 2006*, Toulouse, France