Embodied Model of Speech Production

Juraj Simko

University College Dublin, Ireland

juraj.simko@ucd.ie

Speech production is a well rehearsed, skilled, sequential activity embodied in a complex physical system—our vocal tract. As with any form of skilled action, mastery of speech involves learning how to control and use this system in an *efficient* way within the constraints imposed by functional requirements, i.e. communication.

There are many models which successfully and in great detail address kinematic properties of speech production (e.g. [4, 1, 2]), but fail to take into account the generic principals underlying dynamics of motor organization like energetic efficiency, articulatory ease (possibly the same thing) and communicative efficacy ([3]).

We present an early form of a developing model in which many of the complexities associated with physical modeling of the vocal tract are finessed, while many basic principles relevant to sequencing and coordination in real time with real masses are respected. Our abstract model consists of three simple point-mass pendula hung on massless rods from a common point on a plane. Each pendulum has its own "gravity" force acting on its bob, and these forces act on different pendula from different, evenly spread directions. The pendula are the abstract articulators of our vocal tract model. They do not map directly onto specific articulators, but instead they capture the high-level properties we are interested in. Their movements are subject to physical constraints, and we can specify the dynamics governing their movement in a relatively simple fashion.

Vowel (syllabic nucleus) and consonant production are treated differently in our model. Vowels are represented by shifting equilibrium configurations of the system, while consonants are modeled as ballistic disturbances superimposed over the stream of vowels. Thus, in order to produce a vowel, extra centres of "gravity" can be switched on and off around the system, each acting on one pendulum, forcing pendula to move to a new stable configuration. Consonants are initiated by the brief application of an external forcing impulse on each pendulum separately. Forces with appropriate magnitudes and directions cause pendula to move towards each other until they collide thus reaching an articulatory target ("closure"). Both vocalic equilibria and consonantal collisions represent context-free targets, but the

approach into and path taken from these targets is context sensitive depending on past state and future goals. The model is controlled by four independent *input streams*: vocalic configurations, consonantal force impulses, and modulations of overall system stiffness and damping.

The system is deemed to produce a required segment, if the closeness to target configuration for vowels or closure duration for consonants reaches a specified threshold. These thresholds act not only as minimal requirements for segment production and can be seen as *functional constraints* imposed by the communication environment, but, by lowering and raising them, we can model speaker's control of production on a hypo-/hyper-articulation scale.

As in other biological motor systems, the continuous control parameters of our model offer a potentially infinite set of possibilities for producing an "utterance". The simplicity of the model, however, facilitates the quantification of the expenditure of force used to control the system's articulators for given input streams. This in turn allows us to employ strategies for identifying the input stream constellations which are the *most efficient* solutions for a given sequencing task. As the result, the number of "degrees of freedom" of the system falls dramatically (cf. [3]) and the solutions obtained this way provide an insight into the dynamics of speech production.

A large challenge initially will be to learn to coordinate the input streams so as to elicit speech-like organization and kinematics. Initially, we have the goal of learning to coordinate the consonantal impulses with the stream of vowel configurations. Our aim is to design a developmentally plausible platform for modeling speech acquisition, and implement a neural network control module generating the required input streams. We also intend to use stiffness and damping modulation to model prosodic phenomena such as speech rate, final lengthening, initial strengthening, etc.

References

[1] F. H. Guenther. Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. 102(3):594–621, 1995. DIVA.

[2] I. S. Howard and M. A. Huckvale. Training a vocal tract synthesizer to imitate speech using distal supervised learning. In *Proc. SPECOM 2005*, pages 159–162, Patras, Greece, 2005.

[3] B. Lindblom. Emergent phonology. In *Proc. 25th Annual Meeting of the Berkeley Linguistics Society*, U. California, Berkeley, 1999.

[4] E. Saltzman. The task dynamic model in speech production. In H. F. M. Peters, W. Hulstijn, and C. W. Starkweather, editors, *Speech Motor Control and Stuttering*, chapter 3. Elsevier Science, 1991.