HOW BODIES TALK

Bryan Gick

University of British Columbia
gick@mail.ubc.ca

ABSTRACT

Every speech sound results directly from body movement. However, theories of speech have not always been grounded in biologically plausible theories of how bodies move. High-dimensional models of the body may promise valuable insights into speech, but introduce problems of computational tractability. The present paper outlines some of the phonetic insights we have gained through adopting an embodied framework to enable low-dimensional control of high-dimensional body structures. This framework introduces implications for many aspects of speech sound production, from phonetic universals and the emergence of speech movements to coarticulation and sound change. A key contribution of this work is that it provides a coherent functional unit (the "device") linking biomechanics, perception, production, processing and control of speech sounds. Implications for speech emergence, coarticulation and variation are discussed.

Keywords: Speech production, speech perception, embodiment, biomechanics, motor control.

1. SIMULATING TALKING BODIES

There was a time in phonetics research when talking about speech sounds involved talking about bodies in quite tangible terms. For example, in his early treatment of coarticulation, Martin Joos [1] proposed his 'overlapping innervation wave theory', in which speech movements are controlled by 'innervation waves' of muscle activation. According to this model, coarticulation occurs when these waves of neuromuscular activation overlap. This theory, while interesting, simply proved too difficult to test with the tools available at that time, and was abandoned, leaving subsequent theories less connected to body-based biomechanical and neuromuscular processes. Joos' approach was abandoned not because it was found to lack merit, but rather because at that time the body was far too high-dimensional to model in its full complexity.

In an attempt to fill this long-standing modelling gap, our UBC research group initiated a vocal tract modeling approach led by Sidney Fels which we first presented at the ICPhS 2003 meeting in Barcelona, with the goal of creating a collaborative "extensible infrastructure for a 3D face and vocal-tract model" [2]. This initiative, which was ultimately to become known as ArtiSynth (www.artisynth.org), aimed to create a platform for biomechanical simulation that could help researchers across many fields succeed collectively in modeling the human vocal tract and face. ArtiSynth is now an open-source computational platform for biomechanical modeling with many contributing groups worldwide.

ArtiSynth enables efficient simulation of a large number of connected, dynamic hard and soft tissues of the kind involved in the upper airway. Today, the platform houses hundreds of models used for applications ranging from predictive clinical and surgical modeling and computer animation to biorealistic simulation of speech, swallowing and other airway functions. Currently, the most complete and advanced model in ArtiSynth is FRANK [3], the state of the art in biomechanical modeling of the human head and neck (see Figure 1).

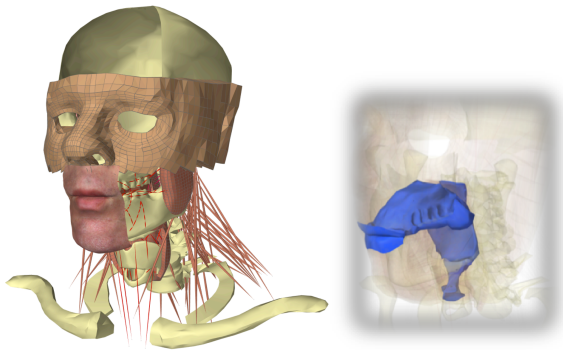

Figure 1: Oblique views of some components of the FRANK model in ArtiSynth, including hard structures and FEM soft tissues (left) and surface mesh of the airway (right).

Though speech was the long-term goal, these models were not created within any specified theoretical framework or with any speech-related constraints or assumptions. Rather, biorealism has been a consistent priority in developing models in ArtiSynth, so that even the earliest component models (e.g, the jaw, tongue, hyoid, etc.) were attempts to simulate those parts of the body in their fullest possible dimensionality. In other words, these models were built not to do speech, but simply to be anatomically correct, based on the best available composite data from medical imaging, high-resolution scanning and fiber-level cryosections. The resulting models can be used not just for speech, but for chewing, swallowing, surgical planning, etc.

2. PROBLEMS CONTROLLING BODIES

Equipped with these high-dimensional models of speech-related body structures, we set out to understand how the body generates the movements that result in speech sound production. However, working with these models presents many of the same challenges as working with real bodies. Early in the evolution of this research, it became clear that the project would quickly run into control problems because of the models' many dimensions, or degrees of freedom (DoF). The term "degrees of freedom" here refers to the number of independent parameters needed to specify the current state of a system.

A human arm, for example, can conservatively be considered as having about 34 degrees of freedom: at least 3 (yaw, pitch and roll) in the shoulder, 2 in the elbow, and 2 in the wrist. Continuing in this manner from one joint to another, the hand adds another 27 or so independent degrees of freedom [4], giving a total of around 34 DoF. Solving a motor task in an unstructured 34-DoF space would mean navigating a possible n^{34} operations, where n is the number of different positions each DoF could take (e.g., the number of different angles at which it is possible to hold the elbow). Thus approached as an unstructured search, controlling a human arm presents a potentially astronomical needle-in-the-haystack problem of computational tractability. This problem compounds when controlling the far more complex vocal tract.

At the outset of this research, our working assumption was that we would be able to use an existing control paradigm such as task dynamics [5] or schema theory [6] to control the various parts of the vocal tract. However, although creating low-dimensional models for motor control is a core goal of these and other motor control approaches, none provide a mapping that is sufficiently detailed to specify control of the fully dimensional speech apparatus. Kelso et al. [7] explicitly avoid providing such a mapping, saying that their approach "is not feasible for the speech articulators whose peripheral biomechanics are much more complex, e.g. […] the tongue and lips" [7, p. 176]. This leaves us with the fundamental problem of identifying the right dimensionality reduction for the high-dimensional speaking human vocal tract. Optimistically, Kelso et al. [7, p. 190] go on to say that while "naturalistic renditions of speech have not told us much (yet) about the speech production process […p]erhaps they will as technology and ingenuity make the speech production system more accessible to observation."

3. MODULARIZING SPEECH

Based on his observations of how multiple joint angles are coordinated in body movements, Bernstein [8] advocated a modular approach to neuromuscular organization as a solution to this problem of dimensionality reduction. This concept of modules refers to coordinated patterns of muscles, sometimes referred to as "muscle synergies". A "module" in this sense unites within a unitary neural structure a set of muscles that, when collectively activated by a single motor command, results in some functional outcome. (see [9, 31]). Applying a modular framework to speech enables us to draw on decades of literature studying the properties of modules in motor control.

Even before Bernstein's work became widely known, speech researchers had discussed essentially modular neuromuscular approaches, such as Cooper et al's. [10] "action patterns," describing speech movements "in terms of a rather limited number of muscle groups" (p. 939). Later, Turvey [11] adopted the concept of "coordinative structures", on which Fowler et al's. [12] speech production model was based. The term "coordinative structure" was coined by Easton [13, p. 591] to describe muscle groupings "underlying all volitionally composed movements, [each] activated by a single command." This term, however, came to be repeatedly redefined in speech circles, making it hard to map them onto bodies.

As early as 1978, Turvey et al. [14: 566] describe coordinative structures as "formally equivalent" to "control space", opening the door to less embodied interpretations. Kelso, Holt, Kugler and Turvey [15] later say the coordinative structure "exhibits behavior qualitatively like that of a force-driven mass-spring system." Subsequent papers describe coordinative structures as "nonlinear oscillators" [16] and as "dynamic patterns" [17]. The definition ultimately settles on "different patterns of articulator cooperation" [18] and "an ensemble of articulators" [7: 29], where "articulators" are abstract task dynamics control structures described by Kelso et al. [5, 7]. To avoid confusion, we generally avoid the term "coordinative structure".

The present paper expands on the original conception of assemblages of nerves and muscles underlying volitional movements. However, the existence of such structures implies the existence of a larger "whole" structure comprising a complete set of dependencies. That is, any active, functioning module in this model inherently constitutes a complete loop that necessarily includes not only neural control and muscle activation, but also the sensory and ecological consequences of the movements generated by those activations, and mechanisms for feedback-based error correction (see Figure 2). While these holistic dependencies are implicit in much of the modular control literature, it is important to acknowledge them explicitly when discussing speech actions, in which the sensory consequences of movements serve not just as feedback to the controller but as the elements of a complex system of communication.

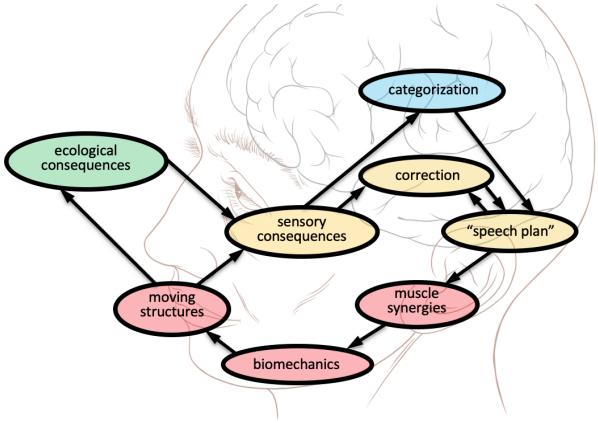

Figure 2: Some aspects of a speech "device".

This holistic view of the module calls to mind the "devices" described by Fowler and Turvey [19], who ascribe to an organism the "capacity to become a variety of special-purpose devices" (p. 11), with each "device" specialized to handle an ecologically relevant task (e.g., a speech sound). A device may thus be thought of as a specific set of dependencies linking modular structures to the ecological functions and feedback loops that drive them. In the context of language, a device is thus a body-based structure that can (minimally) generate a movement and its communicatively relevant (multimodal) sensory consequences. Gick and colleagues [20, 21, etc.] outline some ways a modular framework can generate a wide range of hypotheses for speech.

4. HOW DEVICES CAN INFORM SPEECH

The high degree of realism and complexity in the ArtiSynth models may seem excessive for describing basic speech movements – after all, phoneticians have described speech for centuries with only the most tangential reference to body structures. However, a framework in which the body is controlled using biologically motivated modular structures provides a "loop" of inherent dependencies between perception, processing and production that has long been lacking in speech models. This approach generates a range of testable hypotheses about how speech works. The remainder of this paper outlines a few such properties with implications for phonetics research.

4.1. Robustness

A key property of speech movements is their ability to produce reliable phonetic outcomes despite ever-changing phonetic and non-speech circumstances. E. P. Loeb et al. [22, p. 79] refer to this property as "robustness", observing of limb movement that "there exists a well-defined subset of synergies which will stabilize the limb despite activation noise, muscle fatigue, and other uncertainties – and these synergies stabilize the limb at predictable, restricted locations in the workspace."

In other words, while any set of muscles could in principle act together to generate some output, comparatively few sets will produce consistent, robust outputs. Translating into speech terms, we can assume that there exist a well-defined set of devices that will generate stable, multisensory phonetic outputs (in whatever the relevant communicative space – acoustic, visual, etc.) despite activation noise, muscle fatigue, perturbations from surrounding muscle activations, and other uncertainties.

Robust movements have been associated with properties such as stable cyclicity (as in locomotion [23]) and saturation or "quantality" (see [24, 25]). The property of biomechanical robustness has been observed of universals of speech and emotion expression, including lip rounding/protrusion or closure [26, 27], soft palate configurations [28], and laryngeal states including common phonation types, glottal stop and /h/ [29].

4.2. Emergence and representation

Biomechanically robust structures produce stable links between action and sensory feedback, and are as such more likely to recur and to emerge through use. As G. E. Loeb [30] puts it: "Any adaptive control system will tend gradually toward a locally stable state if one exists." Through frequent use, neuromuscular structure emerges to fit body morphology, biomechanics and ecological function, beginning with repetitive motion in the womb (e.g., [31]). This emergent structure results in natural, functionally determined dimensionality reduction.

Recent approaches to understanding linguistic sound systems have highlighted emergence as a key property (e.g., [32]), often with a focus on the emergence of higher-level phonological and morpho-phonological patterns (e.g., [33]). The present approach enables a body-up model of the emergence of sound systems, allowing us to map our theories of speech directly onto body structures, their biomechanics, kinematics and neural control. This approach predicts that a) similar structures should emerge across languages given roughly similar bodies and ecological functions, and b) emergent structures should bear properties of robustness to internal and external noise sources (e.g., variation in muscle activation, fatigue, phonetic environment, emotion expression, feeding behaviors, etc.).

Of particular relevance to phonetic theory, this model makes no distinction between a device's physical "representation" in the body and its function. That is, each device that emerges from this system has a function. As such, a speech inventory may be thought of as an inventory of body devices, each of which is constructed to generate a particular phonetic event, including all of its internal and external sensory and ecological consequences.

4.3. Temporality, superposition and coarticulation

Devices can be transient, or can operate over an extended time, as in cyclic or tonic activations. Cyclic modules are often associated with locomotion (e.g., [23]), and even these can constrain linguistic actions (as they do when using the forelimbs for sign language movements [34]). Tonic activations, on the other hand, are held across a period of time, as with devices for facial expression or body posture [35]. Speech-related tonic activations include, for example, tongue bracing [36] and pre-/inter-speech posture or articulatory setting [37]. As these kinds of activations are maintained over longer periods of time, they are the also very likely to overlap with activations associated with other speech and non-speech events.

Overduin et al. [38] find that overlapping muscle synergies sum linearly in hand movement control. This property of additivity is reminiscent of Joos' [1] overlapping innervation waves. A simple additive model of this kind has been shown to generate realistic coarticulatory interactions in simulations of overlapping speech movements [39]. Summing activations of speech devices in realistic body models may thus elucidate how coarticulatory interactions may be resolved through low-dimensional physical processes with minimal central control.

4.4. Frequency, complexity and variation

The present framework places no specific limit on the number of solutions that can be generated for a particular task. Indeed, research in posture (see, e.g., [35]) and other areas of motor control indicates that an organism will ideally learn a number of "good enough" solutions for a familiar task [30] rather than a single optimal solution (cf. [40]). Under this view, we should expect that speakers should have learned a range of different variants of each sound to draw on, depending on context and other factors that may affect the speech situation. Such variation has been observed for sounds such as English /r/ [41] and flaps [42], which can alternate even within speaker and phonetic context. The present framework views this not as the exception but as the rule for any movement.

This framework also places no limit on the size or complexity of a device. In general, events that occur with higher frequency are likely to be more redundantly represented. With frequent repetition, even spatially or temporally very complex sequences (such as high-frequency syllables, words or word combinations) may be encoded as their own devices.

For an example of how this plays out in phonology and sound change, consider the case of consonantal weakening. A common conception of speech theories is that a sound may be consistently "weakened", or produced with less effort, apparently causing it to "become" a different sound – a process that is seen as causing languages to change over time. A widely cited instance of this kind of "weakening" is Spanish "spirantization", in which oral closure movements for voiced stops such as /b/ are said to assume a "weakened" form in some contexts, resulting in incomplete closure (e.g, [43, 44]). In the present model, however, a failed /b/ can never properly become a /v/ – at least not by simply changing the degree of muscle activation, since the movements associated with the sounds /b/ and /v/ are produced with distinct sets of muscles rather than a single set of muscles scaled to different activation levels. While it is trivially true that different allophones – even those that may seem on the surface to share important properties such as place of articulation – correspond to categorically different neuromuscular structures, this observation is incompatible with many current approaches to phonology and sound change.

5. CONCLUSION

Phonetic descriptions and speech production models have long imposed traditional dimensionality reductions on the speaking human body. While useful for acoustic synthesis, phonetic description and pedagogy, previous approaches have not provided principled mappings onto real bodies. Approaching dimensionality reduction from the "body up" not only provides us with a principled mapping to bodies, but imposes many constraints on how speech movements might work. A biologically-motivated modular approach opens many novel testable hypotheses that are already helping to shed light on a range of long-standing questions concerning the nature of speech.

6. ACKNOWLEDGEMENTS

Thanks to my many collaborators on this work; NIH Grant DC-002717 and NSERC RGPIN-2015-05099.

7. REFERENCES

[1] Joos, M., 1948. Acoustic phonetics. *Lang.* 24, 5-136.

[2] Vogt, F., Fels, S., Gick, B., Jaeger, C., Wilson, I. 2003. Extensible infrastructure for a 3-dimensional face and vocal-tract model: Proposal for an open source system architecture. *Proc. 15th ICPhS,* Barcelona, 2345-2348.

[3] Anderson, P., Fels, S., Harandi, N.M., Ho, A., Moisik, S., Sanchez, A., Stavness, I., Tang, K. 2017. FRANK: a hybrid 2D biomechanical model of the head and neck. In: Y. Payan, J. Ohayon (eds) *Biomechanics of Living Organs.* Academic Press.

[4] El Koura, G., Singh, K. Handrix: Animating the human hand. *ACM SIGGRAPH 2003*. 110-119.

[5] Kelso, J.A.S., Saltzman, E.L., Tuller, B. 1986. The dynamical perspective on speech production: Data and theory. *J. Phon.* 14(1), 29-59.

[6] Schmidt R.A. 1975. A schema theory of discrete motor skill learning. *Psych. Rev.* 82, 225-260.

[7] Kelso, J., Saltzman, E.L., Tuller, B. 1986. Intentional contents, communicative context, and task dynamics: a reply to the commentators. *J. Phon.* 14(1), 171-196.

[8] Bernstein, N. 1967. *The Coordination and Regulation of Movements*. New York: Pergamon.
[9] d'Avella, A., Giese, M., Ivanenko, Y.P., Schack, T., Flash, T. 2015 Modularity in motor control: from muscle synergies to cognitive action representation. *Front. Compu. Neurosc.* 9, 126.
[10] Cooper, F.S., Liberman, A.M., Harris, K.S., Grubb, P.M., 1958. Some input-output relations observed in experiments on the perception of speech. *2nd Int. Congr. Cybern*. Namur Belg. 930–941.
[11] Turvey, M.T. 1977. Contrasting orientations to the theory of visual information processing. *Psych. Rev.* 84, 67-88.
[12] Fowler, C.A., Rubin, P., Remez, R.E., Turvey, M.T. 1980. Implications for speech production of a general theory of action. In: B. Butterworth (ed) *Language Production*. New York: Academic Press.
[13] Easton, T.A., 1972. On the normal use of reflexes. Am. Sci. 60, 591–599.
[14] Turvey, M.T., Shaw, R., Mace, W.M. 1978. Issues in the Theory of Action: Degrees of Freedom, Coordinative Structures and Coalitions. In J. Requin, *Attention and Performance VII*. Hillsdale, NJ: Lawrence Erlbaum Associates, 557–595.
[15] Kelso, J.A.S., Holt, Kugler, Turvey, M.T. 1980. On the concept of coordinative structures as dissipative structures: II. Empirical lines of convergence. In G.E. Stelmach & J. Requin (ed) *Tutorials in Motor Behavior. North-Holland Publishing Co. 49-70.*
[16] Kelso, J.A.S. 1981. Contrasting perspectives on order and regulation in movement. In J. Long & A. Baddeley (ed) *Attention and Performance IX.* Hillsdale, NJ: Lawrence Erlbaum Associates, 437–457.
[17] Kelso, J.A.S., Tuller, B., Harris, K.S. 1983. A "dynamic pattern" perspective on the control and coordination of movement. In P. MacNeilage (ed) *The Production of Speech*. New York: Springer, 178-173.
[18] Kelso, J.A.S., Tuller, B., Vatikiotis-Bateson, E., Fowler, C.A. 1984. Functionally specific articulatory coopertion following jaw perturbations during speech: Evidence for coordinative structures. *J. Exp. Psych.: Hum. Perc. Perf.* 10, 812-832.
[19] Fowler, C.A.; Turvey, M.T. (1978). Skill acquisition: an event approach with special reference to searching for the optimum of a function of several variables. In: G.E. Stelmach (ed) *Information processing in motor control and learning.* New York: Academic Press.
[20] Gick, B., Stavness, I., 2013. Modularizing speech. *Front. Psychol.* 4, 977.
[21] Gick, B., Schellenberg, M., Stavness, I. Taylor, R. Articulatory Phonetics. 2019. In: W. F. Katz and P. Assmann (eds) *The Routledge Handbook of Phonetics*. New York: Taylor & Francis.
[22] Loeb, E.P., Giszter, S.F., Saltiel, P., Bizzi, E., Mussa-Ivaldi, F.A. 2000. Output units of motor behavior: An experimental and modeling study. *J. Cogn. Neur*. 12, 78-97.
[23] Dominici, N., Ivanenko, Y.P., Cappellini, G., d'Avella, A., Mondi, V., Cicchese, M., Fabiano, A., Silei, T., Di Paolo, A., Giannini, C., Poppele, R.E., Lacquaniti, F., 2011. Locomotor primitives in newborn babies and their development. *Science* 334, 997–999.
[24] Fujimura, O. 1989. Comments on "On the quantal nature of speech" by K. N. Stevens. *J. Phon.* 17, 87–90.
[25] Perkell, J.S. 2012. Movement goals and feedback and feedforward control mechanisms in speech production. *J. Neuroling*. 25, 382-407.
[26] Nazari, M. A., Perrier, P., Chabanas, M., Payan, Y. 2011. Shaping by stiffening: A modeling study for lips. *Mot. Contr.* 15, 141–168.
[27] Gick, B., Stavness, I., and S. S. Fels, C.C., 2011. Categorical variation in lip posture is determined by quantal biomechanical-articulatory relations, in: *Can. Acoust*. pp. 178–179.
[28] Gick, B., Anderson, P., Chen, H., Chiu, C., Kwon, H.B., Stavness, I., Tsou, L., Fels, S., 2014. Speech function of the oropharyngeal isthmus: a modelling study. *Comp. Meth. Biomech. Biomed. Eng. Im. Vis.* 2, 217-222.
[29] Moisik, S., Gick, B., 2017. The quantal larynx: The stable regions of laryngeal biomechanics and implications for speech production. *J. Speech Lang. Hear. Res.* 60(3), 540-560.
[30] Loeb, G. E. 2012. Optimal isn't good enough. *Biol. Cybern*. 106, 757-765.
[31] Keven, N., Akins, K. A., 2017. Neonatal Imitation in Context: Sensory-Motor Development in the Perinatal Period. *Beh. Br. Sci.*, 1-107.
[32] Mielke, J. 2004. *The Emergence of Distinctive Features*. Ph.D. thesis, Ohio State U.
[33] Archangeli, D., Pulleyblank, D. 2015. Phonology without universal grammar. *Front. Psych.* 6, 1229.
[34] Tkachman, O., Purnomo, G., Gick, B. 2018. Cyclic movement primitives underlying two-handed alternating signs in signed languages. *Can. Acoust*.
[35] Ting, L.H., Chiel, H.J., Trumbower, R.D., Allen, J.L., McKay, J.L., Hackney, M.E., Kesar, T.M. 2015. Neuromechanical principles underlying movement modularity and their implications for rehabilitation. *Neuron* 86, 38-54.
[36] Gick, B., Allen, B., Roewer-Despres, F., Stavness, I., 2017. Speaking tongues are actively braced. *J. Speech Lang. Hear. Res*. 60, 494–506.
[37] Gick, B., Wilson, I., Koch, K., Cook, C., 2004. Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica* 61, 220–233.
[38] Overduin, S.A., d'Avella, A., Carmena, J.M., Bizzi, E. 2012. Microstimulation activates a handful of muscle synergies. *Neuron* 76, 1071-1077.
[39] Gick, B., Stavness, I., Chiu, C. 2013. Coarticulation in a whole event model of speech production. *J. Acoust. Soc. Am. – Proc. Meet. Acoust.*, Vol. 19, pp. 060207.
[40] Todorov, E., Jordan, M.I. 2002. Optimal feedback control as a theory of motor coordination. *Nat. Neurosc*. 5, 1226-1235.
[41] Stavness, I., Gick, B., Derrick, D., Fels, S., 2012. Biomechanical modeling of English /r/ variants. *J. Acoust. Soc. Am.* 131, EL355-360.
[42] Derrick, D., Gick, B. 2011. Individual variation in English flaps and taps: A case of categorical phonetics. *Can. J. Ling.* 56(3), 307-319.
[43] Piñeros, C.-E. (2002). Markedness and laziness in Spanish obstruents. *Lingua, 112,* 379-413.
[44] Kaplan, A. (2010). *Phonology shaped by phonetics: The case of intervocalic lenition.* PhD Thesis, UCSC.