Sequencing of Articulatory Gestures using Cost Optimization

Juraj Simko, Fred Cummins

School of Computer Science and Informatics, University College Dublin, Ireland.

juraj.simko@ucd.ie, fred.cummins@ucd.ie

Abstract

Within the framework of articulatory phonology (AP), gestures function as primitives, and their ordering in time is provided by a gestural score. Determining how they should be sequenced in time has been something of a challenge. We modify the task dynamic implementation of AP, by defining tasks to be the desired positions of physically embodied end effectors. This allows us to investigate the optimal sequencing of gestures based on a parametric cost function. Costs evaluated include precision of articulation, articulatory effort, and gesture duration. We find that a simple optimization using these costs results in stable gestural sequences that reproduce several known coarticulatory effects.

Index Terms: articulatory phonology, emergent phonology, dynamics

1. Introduction

Gestures are the atomic units of articulatory phonology (AP) [**?**, **?**, **?**]. They function both as phonological units of contrast, and as phonetically meaningful units of coordination. They have proved to be helpful in understanding a wide range of empirical phenomena, including contextual variation in reduced vowels [**?**], reduction and epenthesis [**?**]. Hitherto, the gestures of AP have been implemented using the task dynamic (TD) framework of Saltzman and Munhall [**?**, **?**], that was originally developed to account for limb movement [**?**]. The theory of AP is formally distinct from its TD implementation. However modeling results that can be compared to empirical data demand some implementation, and TD has served that role exclusively to date.

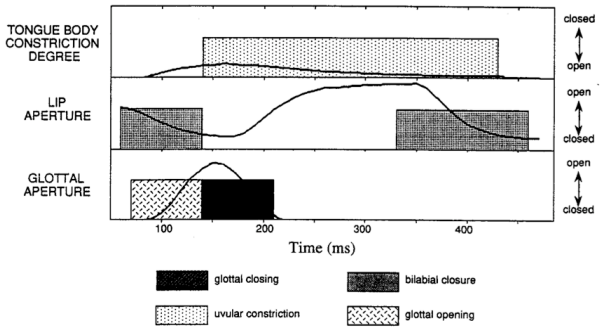

Figure 1: *Gestural score for the utterance* /pʌb/ *and the computed trajectories of the tract variables. From Saltzman (1991).*

Figure 1 illustrates a gestural score for the utterance /pʌb/. Three gestures are shown, each associated with a single tract variable. In general, gestures may comprise more than one tract variable, such as constriction location and constriction degree. A task is then a dynamic specified for a single tract variable. While the gestural score (the filled rectangles) determines when a given gesture is active, the actual tract variable traces are derived from a dynamic implementation, whereby each task is modeled with a simple mass-spring dynamical system. The parameters of the task dynamics are seen as abstract values and do *not* reflect the physical properties of the vocal tract. The mass parameter is arbitrarily set to unity, and the task stiffness is derived so that the resulting tract variable kinematics fits the experimentally established articulatory movement patterns. The system is critically damped, ensuring that when the dynamics are active, the tract variables move smoothly towards their goal states.

From the tract variable trajectories, movements of model articulators are then computed, by mapping from the tract variable space to the space of model articulators. This mapping is purely kinematic; the resulting dynamics imposed on the model articulators cannot be interpreted in terms of forces acting on the individual articulators [**?**]. Whereas tasks and associated tract variables are independent of one another, allowing context free task specification, the set of articulators comprise a common resource, and multiple gestures may interact in influencing any given model articulator. For example, both tongue body constriction and lip opening in the above example will ultimately influence jaw movement, and the resulting movement trace will be a smooth context-sensitive blend of their combined influences [**?**].

2. Sequencing within Articulatory Phonology

The gestural score that provides the starting point for utterance simulation is typically simply assumed [**?**]. The relative timing of the several gestures in a score is, of course, a crucial determinant of the resulting trajectories, and so several attempts have been made to provide principled accounts of their distribution in time, or their phasing with respect to one another. To describe their relative timing, it is usual to treat each activation interval as if it were one cycle of an undamped oscillator, so that points in time can be described as phases. It has been suggested that consonants may tend to be aligned with fixed phases of syllabic nuclei [**?**], but this appears to be somewhat over-restrictive [**?**]. An alternative approach by Byrd is to identify a probabilistic alignment between pairs of gestures, formalized as a window around selected phases [**?**]. Again, however, the selection of the parameters for the alignment windows is somewhat ad hoc. More recent approaches have sought to use a rule based account of coupling relations among oscillators representing individual gestures to derive phasing relationships among gestures in complex speech units and to account for prosodic variation in speech production [**?**]. However, the phasing relations between pairs of

constituent gestures are again set in a pre-determined manner with respect to positional and functional relationships within the given unit (such as onset vs. coda consonant of a syllable) but they do not correspond to the lawful variation of inter-gestural sequencing dependent upon the articulatory nature of embodied gestures [**?**].

3. Embodied Task Dynamics

We have been developing an alternative implementation of task dynamics that differs in one fundamental way from its previous forms. Within our version, tentatively called Embodied Task Dynamics (ETD), tasks are specified as desired locations of *physically instantiated end effectors*, rather than as abstract equilibiria of disembodied dynamical systems. That is, the dynamical parameters of tasks are interpretable in terms of physiological characteristics of the vocal tract end effectors and physical properties of the underlying pure articulators (conceptually equivalent to limb joints of a general motor action system). The task equilibrium point, such as a desired contact between palate and tongue body for a /d/, is expressed within the coordinate space that describes the vocal tract itself. In this case, there are three pure articulators involved in generating the required constriction for /d/ by the tongue tip end effector: the jaw, the tongue body considered relative to the jaw, and the tongue tip considered relative to the tongue body. The stiffness and damping parameters of the task dynamics are derived from the physiologically grounded stiffness and (critical) damping parameters of the pure articulator dynamics as well as from the masses on which the articulators operate. In general, the mapping between tasks (end effectors) and pure articulators may be one to one, one to many, many to one, or many to many. Given the space constraints, an example will serve to illustrate:

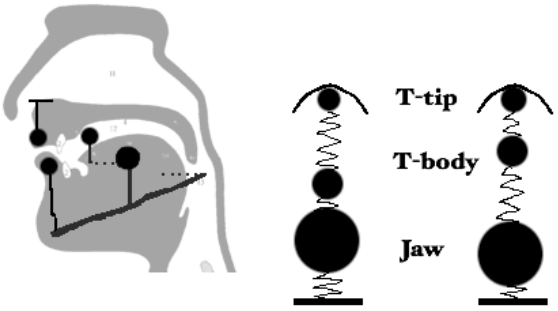

Figure 2: *Jaw, tongue body, tongue tip and lips, as they relate to a model vocal tract, and as implemented. Two behaviorally equivalent articulator configurations are shown.*

Figure 2 illustrates a simple arrangement of model articulators. A jaw, tongue body and tongue tip are free to move up and down (for simplicity, only a single direction of movement is currently considered). The task goal for the consonant /d/ is defined as closure between the tongue tip end effector and the palate. As with conventional TD, this is expressed as a second order dynamic, with resting equilibrium beyond the palate. This single dynamic is transformed into the space of pure articulator movements, yielding motions of the individual articulators. The final trajectories that result will depend, inter alia, upon the individual masses and stiffnesses of the individual articulators. Two articulator configurations are shown on the right hand side. These are equivalent, in that each satisfies the task goal of an alveolar closure, but they demonstrate how the relative contributions of the individual articulators may vary in accomplishing that task goal.

As shown in the left panel of Figure 2, the lips are also implemented. The lower lip is connected to the jaw, while the upper lip has a fixed point of support. As in AP, bilabial task goals are expressed as the distance between the lips, and no independent targets for individual lips are specified.

In contrast to TD where the system's dynamics derives solely from the task level, our model implements a bi-directional mapping between pure articulator and task dynamical parameters; as in TD, the task dynamics imposes coupling among the individual pure articulator dynamics; however, unlike TD, the embodied character of the pure articulators introduces coupling elements among the dynamics of individual tasks. The tasks thus are not mutually independent; their realisation is constrained by physical and anatomical properties of the vocal tract. Moreover, the pure articulators themselves are implemented as final configurations of agonist-antagonist muscle pairs. Space precludes a full description of the implementation. However, our goal here is illustrate how this embodied approach to task dynamics allows the principled derivation of gestural phasing relations. We thus turn now to the issue of optimization based on quantifiable movement costs.

4. Optimization

It is a commonplace within phonetics that observed patterns of articulation result in some sense from a trade off between the conflicting requirements of articulatory ease and the need to be understood. In fact, complex natural systems typically display stable behavior that can be interpreted as approximately optimal with respect to several interacting and conflicting demands. Thus, for example, gait selection in horses may be viewed as optimal with respect to any of several cost functions such as oxygen consumption or musculo-skeletal forces [**?**, **?**]. Interpreting movement as optimal with respect to a parsimonious set of constraints is one way to reduce the apparent complexity in controlling a vast number of degrees of freedom, and to recast movement as coordination within a much more constrained space of optimized action [**?**]. Movement patterns are also the result of optimization processes that operate at many different timescales. Biomechanical constraints evolve slowly over very many generations and reflect a fit between the organism and their typical environment. Over the course of development of an individual, the acquisition of motor skills can also be seen as the discovery of movement forms that are optimal with respect to several sources of constraint, including energy expenditure and behavioral adequacy [**?**].

The implementation of embodied task dynamics allows us to define several candidate cost functions that can be evaluated for any given set of behavioral goals, such as the production of a given sequence, e.g. /abi/[1]. Figure 3 illustrates the optimization procedure we are aiming at. In the top panel, the gestural sequence /abi/ is specified prior to optimization. No phase relations among gestures are presumed. In the bottom panel, gesture activation length and the relative timing of gestures have emerged after the optimization process.

The optimization procedure searches the space of gestural activation patterns and pure articulator stiffnesses for global

[1] In its current implementation, we adopt the expedient of assuming a one-to-one relation between gestures and segments.

minima with respect to a given cost function. That is, we incrementally modify system stiffness and the activation interval onsets and offsets of all gestures until an optimal configuration is found. We assume that our starting constellation (Fig 3, top) is non-optimal with respect to a specific cost function. The cost for the starting configuration is computed, and then the optimization process is employed to perform gradient descent on the cost function, until a local minimum is reached. If the local minimum proves stable with respect to several local perturbations, it is deemed optimal, and this provides us with our final gestural score, and system stiffness.

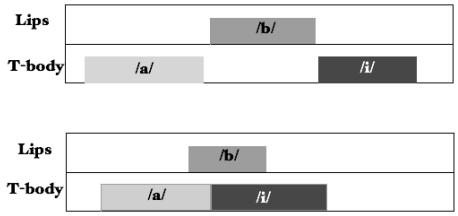

Figure 3: *Gestural score before and after optimization.*

The cost function we currently employ has three components:

$$C = E + \omega_P P + \omega_D D \quad (1)$$

where E is a measure of articulatory effort, P is a measure of communicative efficacy, or parsing cost for the listener, and D is the overall utterance duration. The weights, ω_P and ω_D allow differential weighting of the cost components and are scaled so that the corresponding weight for the effort term, ω_E, has a value of one.

4.1. Articulatory Effort

Articulatory effort, E, is computed based on the evaluation of forces acting at the level of the agonist and antagonist muscles responsible for each individual pure articulator dynamics. Each muscle is modeled as an isotropic elastic spring, so that the overall force depends on the spring lengths and their respective stiffnesses. Stiffnesses, in turn, are always expressed in units of the overall system stiffness that is modified in the optimization procedure along with activation intervals. Our embodied approach to task dynamics allows us to evaluate the realistic forces behind the articulatory movements elicited by the given gestural score.

4.2. Parsing Cost

Minimizing articulatory effort alone would, of course, produce maximally undifferentiated movement and would be useless for communication. The effort constraint thus needs to be balanced by the need for articulatory precision, which in turn will ensure that systematic contrasts are perceived. We therefore penalize movements that are not close to their ultimate targets, and we likewise penalize activations that are too short to be perceived. Together, these considerations motivate a parsing cost, P, which we interpret as related to the ease with which a listener can recover the linguistic contrasts of the utterance. Our simple model currently only allows stop consonants and nuclear vowels. For consonants, we penalize periods during which there is no closure. For vowels, articulatory precision is more gradient in nature, and is indexed by the continuous distance from the intended positional target.

These two cost elements alone, effort and parsing, can be seen as an implementation of Lindblom's Hypo-Hyper continuum [**?**], which assumes that all speech produced represents a trade-off between production costs (articulatory ease) and perception costs (ease of parsing), whereby the ratio, ω_E/ω_P represents the speaker's point along the Hypo-Hyper continuum. Lindblom himself recognized that other constraints are at play, when he observed:

> the assumption about H&H variation being one-dimensional is a deliberate simplification which is likely to be revised in the course of further work. [**?**]

One of the consequences of this simplification is an unrealistic limitation it poses upon speaking rate variation: hypoarticulation must be accompanied by an increased undershoot and shorter duration of realised gestures (faster speech), while hyperarticulation must entail slowing down and a concomitant increase in both precision and in the length of gestural activation intervals. Empirical results, however, do not confirm this direct link between the H&H scale and speaking rate [**?**].

4.3. Duration Cost

The definition of the duration cost, D, associated with an utterance production is straightforward: D is the length of the time interval starting at the onset of the activation interval of the first active gesture in the utterance's gestural score and ending with the realisation of the last gesture, i.e. the closure offset in the case of a consonant and the peak approximation to target in the case of vowel.

Collectively these three costs contribute to the overall cost function, C, which is minimized in the optimization process.

5. Results

One of the first results obtained with the embodied task dynamic model is the simple fact that the optimization procedure converges, and that the resulting movements do not appear to violate any obvious intuitions about gestural sequencing. This may appear trivial, but it needs to be emphasized that the sequences obtained are fully automatic and result from gradient descent based on the above cost function and nothing else. Moreover, by adjusting the duration cost, it is possible to examine gestural sequencing at a range of rates, and in general we find a pronounced phase stability. That is, articulatory phase relations are relatively stable across a range of rates.

Furthermore, our simulations of VCV and VCCV sequences show a clear separation of the vocalic and consonantal tier well documented in phonetic literature [**?**]. In the cost efficient sequences, syllabic nuclei are produced as a continuous sequence of gestural activations interleaved with consonantal gestures. Again, this phenomenon is stable over a range of speaking rates and is an outcome of the cost optimization and is not encoded as an explicit phonological rule. We are now beginning to develop the model to the point where sequence phasing can be compared to appropriate empirical data.

Figure 4 shows optimized gestural scores and associated articulator traces for two utterances: /abi/ and /iba/. The gestural score is shown on top, and the movement traces for the tongue body (solid line) and upper and lower lips (dashed lines). The trace for jaw movement (lighter solid line at the bottom) is also

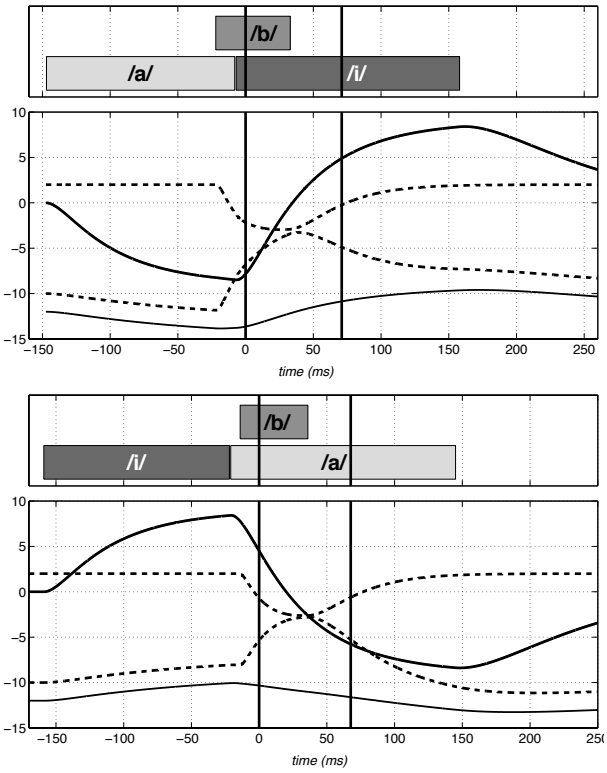

Figure 4: *Articulator traces for utterances* /abi/ *and* /iba/.

shown. The vertical lines demarcate the period of consonantal closure. Lip movement thereafter is due to soft body compression, as the target for each lip is slightly beyond the point at which closure occurs [**?**]. Tongue movement is smooth and continuous during consonantal closure.

For /abi/, tongue movement towards the second vowel is tightly bound to the point consonantal closure, and the onset of lip movement towards the closure starts *before* the tongue movement. This remains robust at a range of rates (not shown), although for faster simulated utterances tongue movement starts relatively earlier during the bilabial gesture activation then for slower ones. On the other hand, for /iba/, lip movements towards the closure starts slightly *later* then tongue movement. These differences in gestural phasing which emerge in our simulations contradict the gesture-independent phasing principles postulated in AP. Interestingly, however, this is precisely the phasing observed by Löfqvist and Gracco [**?**], where 4 of 4 subjects exhibited a slight tongue lead over lips in the sequence /iba/ while three of four exhibited a slight lip lead before tongue movement in the sequence /abi/.

At this stage of model development, a large number of simplifying assumptions have been made in order to ensure that the model remains tractable as the fundamental principles underlying gestural sequencing are developed. Tongue motion is restricted to a single dimension (depicted as vertical above), and only one place target is available for each of tongue tip (alveolar ridge) and tongue body (hard palate). These simplifications allow us to ensure that the procedure for obtaining emergent sequential structure based on physically and psychologically motivated cost terms is correctly implemented. We see no reason in principle why this approach should not extend to a more fully fleshed out vocal tract geometry, and hence lend itself to inclusion within a full articulatory synthesis system.

6. Acknowledgements

This work has been supported by grant number O4/IN3/I568 from Science Foundation Ireland to the second author. Thanks are due to Elliot Saltzman, Louis Goldstein, Dani Byrd and Hosung Nam for extensive discussion, without which this work would remain in its infancy.

7. References

[1] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, pp. 201–251, 1990.

[2] ——, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[3] ——, "Dynamics and articulatory phonology," in *Mind as Motion*, R. F. Port and T. van Gelder, Eds. Cambridge, MA: MIT Press, 1995, ch. 7, pp. 175–193.

[4] ——, "Targetless schwa: an articulatory analysis," *Papers in laboratory phonology II: Gesture, segment, prosody*, pp. 26–56, 1992.

[5] ——, "Tiers in articulatory phonology, with some implications for casual speech," in *Between the Grammar and Physics of Speech: Papers in Laboratory Phonology I.* Cambridge: CUP, 1990, ch. 19, pp. 341–376.

[6] E. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, pp. 333–382, 1989.

[7] E. Saltzman, "The task dynamic model in speech production," in *Speech Motor Control and Stuttering*, H. F. M. Peters, W. Hulstijn, and C. W. Starkweather, Eds. Elsevier Science, 1991, ch. 3.

[8] E. Saltzman and J. A. S. Kelso, "Skilled actions: A task dynamic approach," *Psychological Review*, vol. 94, pp. 84–106, 1987.

[9] D. Byrd, "Influences on articulatory timing in consonant sequences," *Journal of Phonetics*, vol. 24, pp. 209–244, 1996.

[10] ——, "A phase window framework for articulatory timing," *Phonology*, vol. 13, pp. 139–169, 1996.

[11] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proc of the Speech Prosody 2008 Conference*, P. Barbosa and C. Reis, Eds., Campinas, Brasil, 2008.

[12] A. Löfqvist and V. L. Gracco, "Interarticulator programming in VCV sequences: Lip and tongue movements," *Journal of the Acoustical Society of America*, vol. 105, pp. 1864–1876, 1999.

[13] D. F. Hoyt and C. R. Taylor, "Gait and energetics of locomotion in horses," *Nature*, vol. 292, pp. 239 –240, 1981.

[14] C. T. Farley and C. R. Taylor, "A mechanical trigger for the trot-gallop transition in horses," *Science*, vol. 253, pp. 306–308, 1991.

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

[16] E. Thelen, "Time-scale dynamics and the development of an embodied cognition," in *Mind as Motion*, R. F. Port and T. van Gelder, Eds. Cambridge: MIT Press, 1995, pp. 69–100.

[17] B. Lindblom, "Explaining Phonetic Variation: A Sketch of the H&H Theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990, pp. 403–439.

[18] T. Gay, "Mechanisms in the control of speech rate," *Phonetica*, vol. 38, pp. 148–158, 1981.

[19] A. Löfqvist, "Control of oral closure and release in bilabial stop consonants," in *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, P. P. McCormack and A. Russel, Eds. Canberra: The Australian Speech Science and Technology Association, 1996, pp. 561–566.