The Conductor Model of Online Speech Modulation

Fred Cummins

Department of Computer Science,
University College Dublin
fred.cummins@ucd.ie

Abstract. Two observations about prosodic modulation are made. Firstly, many prosodic parameters co-vary when speaking style is altered, and a similar set of variables are affected in particular dysarthrias. Second, there is a need to span the gap between the phenomenologically simple space of intentional speech control and the much higher dimensional space of manifest effects. A novel model of speech production, the Conductor, is proposed which posits a functional subsystem in speech production responsible for the sequencing and modulation of relatively invariant elements. The ways in which the Conductor can modulate these elements are limited, as its domain of variation is hypothesized to be a relatively low-dimensional space. Some known functions of the cortico-striatal circuits are reviewed and are found to be likely candidates for implementation of the Conductor, suggesting that the model may be well grounded in plausible neurophysiology. Other speech production models which consider the role of the basal ganglia are considered, leading to some observations about the inextricable linkage of linguistic and motor elements in speech as actually produced.

1 The Co-Modulation of Some Prosodic Variables

It is a remarkable fact that speakers appear to be able to change many aspects of their speech collectively, and others not at all. I focus here on those prosodic variables which are collectively affected by intentional changes to speaking style, and argue that they are governed by a single modulatory process, the 'Conductor'. In the following section, the Conductor is independently motivated by considering the nature of intentional control of complex action. Converging evidence for the Conductor comes from consideration of the role of cortico-striatal circuits, known to be critical to the online control of action sequences. The Conductor is intended to be a small step towards a biologically plausible account of speech production [23].

In several studies in which subjects mimic other speakers, both Zetterholm [36, 37] and Wretling and Eriksson [35, 14] found that mimics were able to produce reasonable matches to target of such global prosodic parameters as pitch range, global tempo, and phrase size, while many of the fine details of their speech, evident at segmental or subsegmental level, were relatively unchanged,

or changed in ways which were not systematically related to the target. These global prosodic variables may not be independent of one another.

In my own work, I have found that a similar group of variables are affected when two people read aloud at the same time. Speakers in Synchronous Speech experiments have little or no difficulty in reading a prepared text aloud and in synchrony with a co-speaker [9]. In this paradigm, subjects are allowed to read a short text through, and are then given a start signal by the experimenter. Typically, subjects find the task of reading in tight synchrony with another speaker to be a natural one, and their performance is good from the outset, and does not improve much with practice [10]. To satisfy the task goals, they modify their speech rate, inhibit their natural expressive intonation, and produce a rather 'vanilla' form of speech which is, presumably, maximally predictable for their co-speakers. Comparison of speech produced when reading alone and together with another person reveals that there are no clear differences in the relative duration of speech elements across the two conditions [11]. The conditions differ, however, in that temporal variability across speakers is greatly reduced in the synchronous condition for macroscopic intervals, such as phrases and pauses, but unaffected for smaller ones, such as syllables and segments. Pitch variation is also reduced in synchronous speech, and of course the task requirements demand that speakers match their phrase on- and off-sets rather exactly.

A similar bag of variables are communally affected in several motor speech disorders, notably those involving damage to the basal ganglia, as in Parkinson's Disease (PD). The hypokinetic dysarthria typical of this syndrome is characterized (among other things) by difficulty in the initiation of speech, a greatly reduced intonational contour, and altered rhythm, often manifested as rapid but inappropriately modulated syllable sequences [8, 19]. The speech problems experienced by sufferers of PD are clearly related to general motor difficulties, which likewise present as difficulty in initiating action, and disturbed fluency or rhythm once action gets underway.

In a recent thesis, Tyrone [32] argued that dysarthria is a feature of sign language as well as spoken language. Deaf subjects were found to exhibit sign dysarthria in the absence of severe impairment of simple, non-sequenced movements. She concluded that the similarities in vocal and signed dysarthria were rooted in their related demands on the sequencing of complex coordinated movements, rather than in language *per se*. This interpretation receives support from the nature of the difficulties PD patients exhibit in other non-linguistic motor tasks.

One might summarize the variables which are collectively affected by intentional stylistic variation (mimicry, synchronous speech) and by unintentional pathology (hypokinetic dysarthria) as those related to the fluent modulation of speech. Rhythmic modulation, phrase initiation and ending, and intonational variation together make up a set of prosodic variables one might group together under the term of convenience of 'phrasing'. It has been notoriously difficult to cleanly separate linguistic and paralinguistic elements to prosody. As we shall see below, there appears to be considerable overlap in the brain circuits supporting

the modulation of speech in response to a specific speaking situation and those responsible for syntactic sequencing, and so a clean separation of prosody into linguistic and non-linguistic components may not be possible in principle.

2 The Route from Simple Intent to Multiple Effects

If asked to modify one's speech, e.g. by speaking rapidly, or in a very different style, the subjective impression is one of making a relatively simple change. While subjective impressions are not especially trustworthy indicators of mental activity, it is nonetheless striking that a radical change to speech style (yelling, calming voice, comedic variation) is achieved without much conscious detail—one simply shifts from one's regular voice to an altered form—yet the measurable effects are many and varied. In particular, the bag of variables previously grouped under the label 'phrasing' are all going to be affected, yet one does not have the impression of independent variation of each of a host of parameters. (Of course, any given stylistic modification may affect other variables as well, but the phrasing variables identified here are typically affected together.) Rather, these variables collectively characterize specific speaking styles. There is therefore an explanatory gap to be bridged between the subjective experience of a relatively low-dimensional space of intentional speech modification and the observed higher dimensional space of manifest effects.

We have observed that the prosodic variables which collectively constitute the hallmarks of many speaking styles are not independent, but are modulated together. This suggests that a full account of speech production ought to capture their mutual dependence. In what follows, I sketch a preliminary model that does just that. The model draws heavily on an analogy for its initial form, but it will be demonstrated that there is a wealth of neurological evidence, and several relevant and related models, which together suggest that the model is a substantial first step towards a neurobiologically plausible model of online speech production.

3 The Conductor Model of Online Modulation of Speech

The starting point for the present model is an analogy with the conductor of an orchestra. The conductor does not play any instrument herself. In her absence, it may be even possible for an orchestra to get through a musical composition, but the performance will lack the coherence and emotional import of a well-conducted one. The conductor is partly responsible for the sequencing of the individual phrases, but her role most critically affects the temporal and expressive modulation of the individual parts which contribute to the musical whole. Critically, the conductor does not interfere in the high dimensional space of instrument control. Her signals to the individual players are relatively abstract, being restricted to a few dimensions of temporal sequence, relative intensity and their dynamics. (Musical) phrase initiation, cessation and pausing, continuous tempo variation, accentual prominence, are all controlled by the conductor in an

abstract fashion, unencumbered by the differences involved between fingering an oboe and bowing a viola.

One can likewise posit a neurological system which does not, itself, contain detailed instructions for making individual gestures or gesture constellations required for speaking, but which is responsible for sequencing such constellations, and ensuring that they are appropriately modulated, as required by the speaking style employed. The observations made above suggest that this process would affect macroscopic durations, intensity modulation and intonational variation (range, and perhaps accent height). I will refer to this hypothetical process as the 'Conductor'.

In this view of speech production, elements are retrieved from some source, and are sequenced and modulated during online production. The retrieved elements themselves contain the gesture-specific information required for production. The conductor is responsible for the temporal sequencing of these gestures, including the responsibility for ensuring that such sequencing is fluent and context sensitive. The conductor is also responsible for the affective modulation of the units sequenced, that is, intensity and pitch modulation which is not specified by the concatenated units, but is a function of the specific communicative situation, including speaker, and listener-oriented constraints. This modulation is relativly abstract, and may be thought of as akin in some respects to continuous variation along the hypo-hyper axis of variation, as in Lindblom's H&H theory of speech production [25].

The model is agnostic about the exact nature of the elements sequenced, but we note that they can hardly be much larger than syllables, or much smaller than segments. The collection of gestures which are phased with respect to the syllable nucleus in Articulatory Phonology provides a plausible candidate unit size [5, 6] which may serve for initial development of the model. (As an aside, it is interesting to ask what size the units sequenced by a conductor are, or, indeed, whether the question is meaningful.)

4 Implementing the Conductor Within a Production Model

The framework of Articulatory Phonology (AP), and its implementation using Task Dynamics, provides an initial insight into how the Conductor might operate during production. Some recent work within AP has sought to incorporate abstract gestures, which are, themselves, not tied to specific articulators. A 'prosodic-gesture' or 'π-gesture' is employed to modulate the temporal unfolding of a group of physical gestures which are linked to specific articulators [7]. The AP model allows at least two distinct modulation options here: the stiffness of individual gestures, or the clock-rate which underlies the dynamics of all gestures. Although current opinion seems to favour the latter as a modulation mechanism (see also [28]) it is probably too early to be dogmatic about the exact method of modulation employed. Modulation of these parameters alone brings with it changes to the relative alignment (and hence the fluent context-conditioned se-

quencing) of elements, and also has consequences for the extent of the resultant gestures, as demonstrated in Byrd and Saltzman (2003). This model thus provides a natural framework for the future development of the Conductor model.

Articulatory Phonology is not the only framework in which a process akin to the Conductor could be implemented. It is also the case that the adoption of the AP framework requires a commitment to several choices which are not necessary elements of the Conductor model. For example, the proposed π-gestures in AP are constrained to affect all concurrent gestures similarly. This is perhaps a sensible requirement, but it is not a necessary consequence of the Conductor model. In addition, the simple second-order dynamic associated with individual gestures within AP constrains the number of possible variables which could be affected by the Conductor, effectively limiting them to stiffness modulation and time warping. Other approaches to gestural modelling might provide a different set of potential implementation variables, and each such set will impose different limits on the effects which a relatively abstract and non-specific Conductor can bring about. But it is a strong contention of the present model that any plausible account of speech production must allow this kind of abstract, gesture-independent modification by an external process.

One clear responsibility of the Conductor is the regulation of speaking tempo. It has been repeatedly observed that the bulk of tempo variation in speech production is effected by adjusting the duration and relative frequency of pauses [31]. That is, it is the initiation and cessation of individual phrases which underlies most of the perceived tempo of speech, not any direct modification of the internal details of segments or syllables. This is clearly compatible with a Conductor process whose primary task is the fluent sequencing of relatively invariant units in production. Further tempo modulation, corresponding to changes in articulation rate, can be achieved by varying the above stiffness and clock variables.

5 A Neurophysiological Basis for the Conductor

I propose that there is, in fact, a neurophysiological system which implements the Conductor process. The proposed role of the Conductor seems entirely compatible with our current knowledge of the role of specific circuits originating in motor and pre-motor cortex, extending by a variety of routes through the basal ganglia, onwards through the thalamus and back to cortex. There are several parallel circuits known to exist, and they include both direct and indirect paths through the basal ganglia [12, 17, 3].

Proposed functions of these cortico-striatal loops, as they are sometimes called, include the selection of some actions and inhibition of others, the rule-based sequencing of actions, and the coordination of action sequences into fluent wholes [3]. The role of these circuits in speech production has been the focus of some empirical and modelling work [23, 33], as discussed below.

Connectivity between the individual stages of the cortico-striatal loops suggests a funnelling of information, or dimensionality reduction between the cerebral cortex and the basal ganglia. A recent model, the Reinforcement Driven

Dimensionality Reduction model of Bar-Gad, Bergman and colleagues has made the postulate explicit that the basal ganglia are compressing cortical information using optimal information extraction methods [2, 3]. This dimensionality reduction which appears to take place suggests that if the basal ganglia are modulating the sequencing and execution of individual components, they are doing so in a lower dimensional space than that which specifies the execution of each individual component. In short, the funneling taking place at the basal ganglia appears to be *prima faciae* suited to implementing a low-dimensional control signal which modulates the individual motoric components in relatively non-specific fashion, as envisioned by the Conductor model.

The issue of low-dimensional control over a complex, high-dimensional system addresses both issues raised at the outset. It accords well with the intuition that control is relatively abstract and goal-directed, and does not involve detailed and disjoint control over the myriad of variables affected by a change in style. This is a solution which addresses the infamous 'degrees of freedom' problem noted by Bernstein [4], and is similar to action-theoretic approaches to skilled action, in which task-specific goals are defined in a relatively low-dimensional space, and they cause multiple, mutually yoked effects in effector space [20, 21]. It also follows that low-dimensional control of a higher-dimensional system will have, as a necessary consequence, the co-variation of very many variables in the more complex system.

A caveat is in order, before the hypothesized Conductor is identified with specific neural circuits. Although the cortico-striatal loops are clearly implicated in rule-based sequencing, context-conditioned action modulation and fluency, all of which suggest a pivotal role in speech production, there are several such circuits, which may differ greatly in their relative contributions, and the parallel circuits may not be entirely separate. The circuits are also implicated in other, rather distinct activities, such as reward-based action. Matsumoto and colleagues [27] have shown that CS-loops may be essential to the *acquisition* of smooth movement patterns, but that they may not be essential to their *execution*, though this evidence is based on two primates only. And similar circuits linking cortex, thalamus and the cerebellum are also regularly implicated in the fluid control of action. Indeed the cerebellar loops may jointly regulate fluent action sequencing in tandem with the cortico-striatal loops [30]. The insula has also been implicated in the coordination of speech articulation [13].

6 Relation to Other Models of Speech Production

The view of speech production sketched herein suggests some answers to rather fundamental issues in modelling speech production. It is assumed that relatively context-insensitive representations are available for sequencing and for context-specific modulation by the Conductor. This rather weak claim accords with most views of the process of speech production, and is thus relatively uncontroversial. However, the Conductor models assumes that those forms which are available for sequencing are already specified in a form suitable for online modulation

by the Conductor. I have suggested that the syllable representations employed within Articulatory Phonology suggest themselves as possible units. One reason this is so is that a gestural specification of linguistic forms provides some rather obvious channels for modulation to be effected, via time warping or stiffness change. More conventional phonological representations which employ timeless, abstract symbols pose huge problems of translation into some form suitable for production [16].

The proposed Conductor model is not at all incompatible with some existing models of speech production. Firstly, the abstract process of modulation through time warping or stiffness variation suggested here is one possible way of implementing Lindblom's continuum of Hyper- and Hypo-speech [25]. The H&H model emphasizes the fact that speech production is adaptive and finely modulated, so as to respect both speaker and listener-oriented constraints. The modulation envisaged within this model is responsible for a myriad of kinematic effects, but these are understood to derive from a much simpler, low-dimensional control space.

Lieberman [23, 24] has developed a theory of the evolution of speech, in which speech production is based around what he calls "Functional Language Systems", implemented by distributed networks within the brain. The cortico-striatal loops which are implicated by the Conductor model are here hypothesized to underlie sequencing of both speech/motor elements and syntax.

Ullman [33, 34] has developed a model in which declarative and procedural elements are fundamentally separated. The declarative elements correspond roughly to lexical units, while the procedural systems are responsible for their sequencing. He explicitly identifies the cortico-striatal circuits, along with the cortico-cerebellar circuits previously mentioned, as supporting the procedural system. The sequencing of elements treated in Ullman's model refers to syntax, rather than the 'phonetic' sequencing discussed above. Indeed, there is good reason to think that the sequencing abilities of the cortico-striatal circuits might serve both purposes: the physical stringing together of units into a fluid sequence of sounds, and the rule-based serial ordering of units retrieved from the lexicon, and ordered in accordance with the rules of a grammar. Some general notes on sequencing now follow.

7 On Sequencing

The basal ganglia and associated circuits are phylogenetically old, going back at least to the common ancestor of the human and the frog. In rats, cortico-striatal loops (CS-loops) are critically implicated in grooming behaviour where such grooming consists of syntactically well-formed sequences of highly practiced actions [1]. Neuronal activity in the CS-loops is not a function of the individual movements (which may occur within our outside of syntactically governed sequences), but of the syntactic sequence itself. Rat grooming syntax is not hierarchically complex, but involves sequencing of specific action types. Graybiel [18] has argued that one role of the basal ganglia in sequence learning includes

the recoding of action elements into higher-order units: a form of 'chunking' for action. Fentress [15] has demonstrated that the acquisition of the adult grooming pattern in mice is more than just learning to string appropriate movements together. Baby mice learn individual grooming strokes in isolation, and then have to learn to integrate them within a fluent action sequence. Initial attempts to generate a fluent sequence appear to result in a temporary 'unlearning' of the individual parts, as the sequencing itself is mastered. This again points to a clean separation of the problem of fluent sequencing from that of the execution of individual actions in isolation. It also suggests that part of what is being mastered is the hierarchical organization of action sequences, and not just a linear ordering. The hierarchical nature of sequential action in humans has also been demonstrated by Rosenbaum [29].

It may appear as if two entirely separate roles for the CS-loops are being sugggested. On the one hand, they are clearly implicated in syntactic sequencing. This is the domain of formal linguistics, and is typically considered to be entirely disjoint from the messier business of producing sounds. On the other, the same circuits are suggested to be responsible for the fluent production of context-conditioned speech.

Perhaps the separation of disciplines typically enshrined in our academic departments and professional societies may not adequately reflect the partition of labour as embodied in real brains [23, 24]. If language is not to be considered as miraculous, it must indeed be based on cognitive abilities which precede it phylogenetically. The close association of syntactic sequencing and the fluent sequencing of complex skilled action was famously pointed out by Karl Lashley [22]. The convergence of behavioural and neurophysiological evidence sketched above seems to suggest that we may be within sight of an account of language which is credible diachronically in evolutionary terms, and synchronically in neurophysiological terms. Despite the difficulties this may pose for trade unions in Universities, it is surely to be welcomed.

References

1. J. Wayne Aldridge, Kent C. Berridge, Mark Herman, and Lee Zimmer. Neuronal coding of serial order: syntax of grooming in the neostriatum. *Psychological Science*, 4(6):391–395, 1993.
2. I. Bar-Gad, G. Havazelet-Heimer, J. A. Goldberg, E. Rupping, and H. Bergman. Reinforcement-driven dimensionality reduction—a model for information processing in the basal ganglia. *J. Basic and Clin. Physiol. Pharm.*, 11(4):305–320, 2000.
3. I. Bar-Gad, G. Morris, and H. Bergman. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71:439–473, 2003.
4. N. Bernstein. *The Coordination and Regulation of Movements*. Pergamon Press, London, 1967.
5. Catherine Browman and Louis Goldstein. Some notes on syllable structure in articulatory phonology. In Osamu Fujimura, editor, *Articulatory Organization: Phonology in Speech Perception*. S. Karger, Basel, 1988.

6. Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.
7. Dani Byrd and Elliot Saltzman. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2):149–180, 2003.
8. G. J. Canter. Speech characteristics of patients with parkinson's disease: I. intensity, pitch, and duration. *Journal of Speech and Hearing Disorders*, 28(3):221–229, 1963.
9. Fred Cummins. On synchronous speech. *Acoustic Research Letters Online*, 3(1):7–11, 2002.
10. Fred Cummins. Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31(2):139–148, 2003.
11. Fred Cummins. Synchronization among speakers reduces macroscopic temporal variability. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 304–309, 2004.
12. M. R. DeLong. Overview of basal ganglia function. In Mano et al. [26], pages 65–70.
13. Nina F. Dronkers. A new brain region for coordinationg speech articulation. *Nature*, 384:159–161, 1996.
14. Anders Eriksson and Pär Wretling. How flexible is the human voice?–a case study of mimicry. In *Proceedings of EUROSPEECH*, volume 2, pages 1043–1046, Rhodes, Greece, 1997.
15. John C. Fentress. Hierarchical motor control. In *Psychobiology of Language*, pages 40–61. MIT Press, 1983.
16. Carol A. Fowler, Philip Rubin, Robert Remez, and Michael Turvey. Implications for speech production of a general theory of action. In B. Butterworth, editor, *Language Production*, pages 373–420. Academic Press, San Diego, CA, 1981.
17. Ann M. Graybiel. The basal ganglia. *Trends in Neuroscience*, 18(2):60–62, 1995.
18. Ann M. Graybiel. The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, 70:119–136, 1998.
19. L. Hammen, Vicki and Kathryn M. Yorkston. Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria. *Journal of Communication Disorders*, 29:429–445, 1996.
20. Katherine S. Harris. Action theory as a description of the speech process. In Herman F. M. Peters and Wouter Hulstijn, editors, *Speech Motor Dynamics in Stuttering*, chapter 2, pages 25–39. Springer, New York, 1987.
21. J. A. S. Kelso, K. G. Holt, P. N. Kugler, and M.T. Turvey. On the concept of coordinative structures as dissipative structures: II. Empirical lines of convergence. In G.E. Stelmach and J. Requin, editors, *Tutorials in Motor Behavior*. North-Holland, 1980.
22. Karl S. Lashley. The problem of serial order in behavior. In L. A. Jefress, editor, *Cerebral Mechanisms in Behavior*, pages 112–136. John Wiley and Sons, New York, NY, 1951.
23. Philip Lieberman. *Human Language and Our Reptilian Brain: The Subcortical Bases of Speech, Syntax, and Thought*. Harvard University Press, 2000.
24. Philip Lieberman. On the nature and evolution of the neural bases of human language. *Yearbook of Physical Anthropology*, 45:36–62, 2002.
25. Björn Lindblom. Explaining phonetic variation: a sketch of the H&H theory. In William J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic, Dordrecht, 1990.
26. N. Mano, I. Hamada, and M. R. DeLong, editors. *Role of the Cerebellum and Basal Ganglia in Voluntary Movement*. Elsevier, 1993.

27. N. Matsumoto, T. Hanakawa, S. Maki, A. M. Graybiel, and M. Kimura. Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *Journal of Neurophysiology*, 82:978–998, 1999.
28. Robert Port and Fred Cummins. The English voicing contrast as velocity perturbation. In J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe, editors, *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 1311–1314. University of Alberta, 1992.
29. David A. Rosenbaum, Sandra B. Kenny, and Marcia A. Derr. Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 9(1):86–102, 1983.
30. W. T. Thatch, J. W. Mink, H. P. Goodkin, and J. G. Keating. Combining versus gating motor programs: differential roles for cerebellum and basal ganglia. In Mano et al. [26], pages 235–245.
31. Jürgen Trouvain. *Tempo Variation in Speech Production*. PhD thesis, Institut für Phonetik, Universität des Saarlandes, 2004. Published as Forschungsbericht Nr. 8.
32. Martha Ellen Tyrone. *An Investigation of Sign Dysarthria*. PhD thesis, The City University, London, 2004.
33. M. T. Ullman, S. Corkin, M. Coppola, G. Hickok, J. H. Growdon, and W. J. Koroshetz. A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are procesed by teh procedural system. *J. Cognitive Neuroscience*, 9(2):266–276, 1997.
34. Michael T. Ullman. Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, 92:231–270, 2004.
35. Pär Wretling and Anders Eriksson. Is articulatory timing speaker specific? – evidence from imitated voices. In *Proc. FONETIK 98*, pages 48–52, 1998.
36. Elizabeth Zetterholm. Intonation pattern and duration differences in imitated speech. In *Proc. Speech Prosody 2002*, pages 731–734, Aix-en-Provence, 2002.
37. Elizabeth Zetterholm. The same but different: three impersonators imitate the same target voices. In *Proc of 15th ICPHS*, Barcelona, 2003.