PLANART: A MODULAR PLATFORM FOR COMPUTATIONAL MODELING OF ARTICULATORY PLANNING

Alice Turk[1], Benjamin Elie[1], Juraj Šimko[2]

[1]The University of Edinburgh, [2]University of Helsinki
a.turk@ed.ac.uk, benjamin.elie@ed.ac.uk, juraj.simko@helsinki.fi

ABSTRACT

We introduce the initial developments of PlanArt, a modular platform for testing different computational modeling approaches for speech articulation planning. The impetus for building this software comes from theoretical debates in the fields of Phonology and Phonetics which have led to vastly different modelling architectures, proposed mechanisms, as well fundamentally different dynamic models of speech articulatory trajectories. One way to address these debates is to test different models using a common platform for model testing in different contexts. In this paper, we present the platform architecture, designed to initially compare two Optimal Control Theory approaches: (1) Šimko & Cummins' Embodied Task Dynamics based on Articulatory Phonology, and (2) Turk & Shattuck-Hufnagel's XT/3C approach based on symbolic phonological planning. In addition, we discuss some of the challenges of this type of modelling task.

Keywords: speech articulation, articulatory modeling, speech production, planning, phonetics and phonology

1. INTRODUCTION

The creation of PlanArt is inspired by debates in the theoretical literature about the nature of representations and processes in planning speech articulations. Our hope is to contribute to this debate by providing a single modular platform for the comparison of models used for predicting articulatory patterns in different segmental and prosodic contexts. These comparisons will allow us to better understand different computational approaches to speech production planning as well as identify the interdependencies inherent within each approach. We believe that the comparative platform will ultimately lead to a fair evaluation of different theoretical approaches to speech motor control.

Here we first introduce the theoretical issues that motivate the development of PlanArt, followed by the goals of our system and the general architecture of the modular platform. We will illustrate the platform by the current partial implementation designed to compare the architectures and components required for two existing approaches to speech articulation modelling:

(1) **ETD**: A development of Articulatory Phonology called Embodied Task Dynamics [1, 2], and

(2) **XT/3C**: Turk & Shattuck-Hufnagel's Phonology-extrinsic Timing/3-Component approach [3, 4].

These two approaches both subscribe to the principles of Optimality Control Theory (OCT) which assume that speech articulation patterns arise as an optimal solution to both production and perceptual constraints (see Section 3).

Some of the differences between the two approaches that will be addressed by the PlanArt platform include (1) a symbolic planning component in XT/3C that is not used in ETD; (2) speech production goals defined in terms of acoustic cues in XT/3C *vs.* articulatory constriction goals in ETD, and (3) different dynamic models for trajectory planning: second order dynamics for ETD and Lee's General Tau Theory control [5] for XT/3C.

2. THEORETICAL BACKGROUND

This section provides a general discussion of different theoretical issues and their implications for computational models of speech articulation, to be addressed by a complete version of PlanArt.

2.1. Systematic phonetic variability

Words that speakers consider 'the same' (*phonologically equivalent*) are nevertheless pronounced systematically differently in different contexts (often referred to as *systematic phonetic variability*). Although there is strong agreement that all words that sound the same must share sub-lexical phonological representations, there is lack of agreement about the nature of these representations (spatiotemporal vs. symbolic). These different types of proposed representations lead to different system architectures that relate these representations to produced speech that

includes appropriate variability in surface articulatory patterns. Models assuming symbolic phonological representations traditionally separate phonological and phonetic planning in the speech production process and require mapping the Phonological (symbolic) representations to the Phonetic (articulatory) space [3].

An alternative approach assuming a spatiotemporal (gestural) nature of phonological units [6, 1] provides a way of unifying the planning process into a single Phonological planning component.

While there is growing appreciation for the influence of multiple phonological and stylistic factors on the phonetic form of particular utterances, there are different ways of modelling this, e.g. Optimal Control Theory (e.g. [7] and Dynamic Field theory [8]).

2.2. The underlying dynamics of speech movements and the nature of speech goals

Most speech movements exhibit the hallmarks of practiced, purposeful movement, that is, single-peaked, and (often but not always) symmetrical velocity profiles. Modifications of damped oscillator-based systems have been previously used to generate speech movement trajectories in several articulatory models [9, 10, 11]. Lee's General Tau theory [5] (see also below) addresses the issue of targeted motor action more generally. It controls the time-course of movement (movement acceleration and deceleration over time) via a single parameter in a relatively simple equation. Most importantly, this equation presents a better fit to articulatory measurements than oscillator-based systems [12].

A closely related issue is the nature of the goals of speech movements. Models with context-independent articulatory goals (such as most damped oscillator-based ones) require accounts of contextual differences in spatial positions at movement endpoints, e.g. via undershoot from shorter activation intervals, gestural blending, or sensitivity to feedback about target achievement. Models with context-dependent goals–such as those using General Tau theory that require full spatial and temporal specification of the realized movement targets–require ways of determining these goals.

While models with articulatory goals (where sound is the result of motor action and does not need to be explicitly specified) are arguably simpler, acoustic goals may be required to account for the equivalence of different motor actions that result in similar sounds [13]. Approaches relying on acoustic speech targets need a mechanism for solving the acoustic-to-articulation inversion problem [14].

2.3. The nature of coordination and speech timing patterns

Speech involves the skilled, efficient coordination of sets of articulators. However, how coordination works is still unclear. Possibilities include coordination based on the time of movement onset, the time of target achievement, and based on feedback about particular spatial or spectral states (e.g. associated with target achievement). Different dynamic models of speech movement trajectories have implications for modelling coordination. Because targets are never reached in damped oscillator-based systems, these appear to require either coordination based on movement onsets [15], or coordination based on feedback about target achievement in articulatory or acoustic space [14, 16, 17]. In contrast, Lee's Tau Theory equation provides a mechanism for coordination based on the *time* of target achievement.

Debates relate to whether speech timing patterns emerge from adjustments of default activation intervals of spatiotemporal phonological representations [6], from sensitivity to feedback about when speech targets are achieved [16, 17], or if they are explicitly specified, as part of a Phonetic Planning process [3], via optimization procedures [2], [1].

3. MODELING OBJECTIVES

Our overall goal is to develop a flexible and modular modelling platform allowing for parallel implementation, testing and comparison of sets of competing theoretical comparisons, such as those introduced above.

In the initial stages of development, our focus is on creating a platform architecture that allows a comparison of two theories–ETD and XT/3C–that share a common Optimal Control Theory approach in which trade-offs between competing requirements of production efficiency and perceptual efficacy are satisfied [18, 3].

The modularity of the platform assumes that the transformation of the input to the model–expressed in a form of a gestural score (for ETD) or a sequence of symbolic phonemic units with associated acoustic cues, for XT/3C–to the output (articulatory trajectories) proceeds in identifiable discrete steps, and that each step can be computationally implemented as an independent module with its own inputs and outputs. The platform will be flexible in the sense that individual modules can be replaced by other modules with essentially the same functionality albeit using different theoretical underpinnings.

Briefly, the competing requirements are concep-

tualized in the form of a multi-objective cost function, combining an articulatory *effort* E, linked to the forces applied to embodied speech articulators during speech, requirements of perceptual distinctiveness expressed as a *parsing* cost, denoted P, and the cost of *time* it takes to produced the entire utterance, denoted D. The composite cost function can then be expressed as a weighted average of these partial cost components

$$C = \alpha_E E + \alpha_P P + \alpha_D D, \quad (1)$$

where α_E, α_P, α_D are weights assigned to individual components within the trade-off.

The weights of these components and, potentially, other parameters of the cost function (*optimization parameters*) quantify the relative priorities of task requirements, including context-dependent constraints, as well as costs of movement (e.g. effort). The task requirements are assumed to be provided to the optimization procedure (conceptualized as part of *dynamic planning*) through prior phonological processing. For a given intended utterance, the optimal timing and kinematic characteristics of speech movements that minimize the value of the combined cost function reflect the requirements encoded by the optimization parameters. Increasing the weight α_D, for example, imposes a greater premium (cost) on time, and the optimal trajectories can be expected to correspond to articulation at a faster speaking rate; this type of adjustment can be imposed globally or locally, eliciting global or local articulation rate changes.

4. PLATFORM ARCHITECTURE

The modular PlanArt architecture is illustrated in Fig. 1. The Symbolic Planning Component, required only for XT/3C (corresponding to its Phonological Planning Component) provides the parameters for the task requirements (context-dependent constraints) to be used in the objective function.

The Dynamic Planning module is the component implementing the optimization procedure discussed above. This processing step leads to generation of a detailed articulatory plan in the form of a vector of static and dynamic articulatory parameters that can be executed during a speech production event by the Motor-Sensory Implementation module. During the optimization process itself, the Articulatory trajectory module (potentially a copy of the Motor-Sensory Implementation one) computes the articulatory trajectories (given a static and a dynamic model) and returns a cost C to the dynamic plan following Eq. (1). The optimization procedure consists of finding the articulatory plan for which the

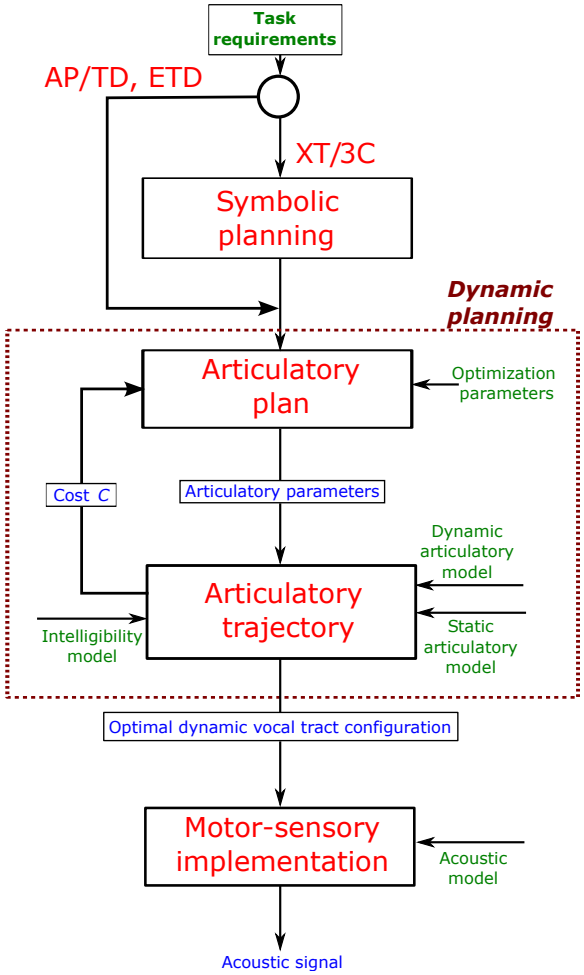

Figure 1: Architecture of the PlanArt software for comparing ETD and XT/3C. Elements in green and elements in blue are module inputs and outputs, respectively. Modules names are in red fonts.

articulatory trajectory component returns a minimal cost. Within ETD this thus corresponds to a gestural score formed during phonological planning (cf. Articulatory Phonology), where the gestural score is a phonological representation of an utterance. The dynamical planning stage corresponds to the Phonetic Planning Component within XT/3C theory. The optimal articulatory trajectories are then fed to the motor-sensory implementation which computes the acoustic signal given a user-defined acoustic model.

The modular architecture can also capture other key differences between the two approaches, such as different dynamic models for trajectory planning.

As a development of the Task Dynamics implementation of AP [19, 20], ETD uses a (mutually coupled system of) second order critically damped dynamics to model the time-course of articulatory movement and task realization. While this choice does not yield the best fitting articulatory movements, it reflects the theoretical commitment of the

approach, namely seeing the task oriented articulatory action (a gesture) as a phonological primitive. As in Articulatory Phonology, systematic phonetic variability results from the properties of the dynamics. For example, the *phonological* target conceptualized as a parameter of dynamical description is never reached, and variability in ETD arises partly from a degree of undershooting the target. The degree of undershoot in turn depends on other dynamical parameters, such as stiffness and activation interval, that are within ETD computed as a part of the phonological planning step as optimal solutions of the objective function.

On the other hand, XT/3C proposes the time between acoustic landmarks, as well as General Tau Theory and its equations of movement as underlying sources of timing and articulatory dynamics which are used in its Phonetic Planning Component. Unlike damped mass-spring dynamics, the Tau Guide equation assumes a pre-specified movement duration and a *reached* target of the movement.

Another difference between the XT/3C and ETD approaches is related to speech production goals defined in terms of acoustic cues for the former *vs.* articulatory constriction goals for the latter. Within the Optimal Control driven PlanArt architecture, the appropriate degree of proximity to the target is evaluated within the parsing cost part of the objective function. Consequently, this modeling decision (acoustic or articulatory targets) can be to a large degree implemented as part of the mathematical definition of the cost component and does not necessarily influence the remaining parts of the implementation.

In this way (and leaving aside some theoretical assumptions), the modular architecture of PlanArt platform will thus allow testing various setups by, for example, swapping the dynamic models used for trajectory planning or the nature of targets between two approaches. This possibility will allow us to test the practical relevance and importance of various modelling commitments (although see some potential challenges to the modular approach below).

5. MODELING CHALLENGES

The optimization approach facilitates the modular architecture of the PlanArt system by allowing inclusion of alternatives in terms of dynamic models of articulation (tau-guided *vs.* damped attractor dynamics), or in terms of the nature of targets. The details of the implementation might, however, differ quite dramatically depending on the choice, in particular in terms of the overall number of optimized parameters.

The AP approach (including ETD) assumes context-independent phonological targets (in the task space) implemented as target parameters of the dynamical description; each possible gesture within the given language repertoire has one such target. The contextual variation arises through the context-dependent nature of realization of the movement towards this target, determined by other dynamical parameters (such as stiffness) and by the activation patterns. For a given utterance, the parameters to be optimized are thus limited to this relatively small "context dependent" subset.

On the other hand, an assumption of context dependent targets (articulatory or acoustic) implies a need to optimize much higher number of parameters. For a Tau-guided movement, for example, every articulatory movement has to be fully specified in terms of its (achieved) end-point and duration of the movement. That means that in addition to the dynamical parameters of the context-dependent movement dynamics (a Tau-coupling constant in case of the Tau-guided movement), the optimization process needs to find appropriate context-dependent targets for each individual articulator for each allophonic token within the utterance.

As argued above, thanks to the optimization approach this is not a problem in principle, but may have practical (and, possibly, theoretical) implications regarding the potential efficacy of the system.

Another modelling challenge will be the appropriate implementation of acoustic (rather that articulatory) targets of speech production that will essentially require an explicit inverse mapping from acoustic to articulatory space. The team developing the PlanArt software are currently exploring probabilistic models specially designed for this task.

6. DISCUSSION

Model comparison via a common platform is a promising way of assessing the benefits and computational challenges of theoretical alternatives. Our hope is that PlanArt can be used by the research community to compare different modeling approaches reflecting different theoretical assumptions (beyond the presently discussed ETD and XT/3C). The software is intended to be an open-source software. For the sake of anonymity, the link to a public repository, where the source codes of the software are freely available, will be provided only in the final version of the paper. PlanArt is in its beginning stages; we anticipate many insights will come of this project.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the European Research Council (ERC) under the European Unionâs Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; ERC Advanced Grant awarded to A. Turk; Grant agreement No. 101019847).

8. REFERENCES

[1] J. Simko and F. Cummins, “Embodied task dynamics,” *Psychological review*, vol. 117, no. 4, p. 1229, 2010.

[2] ——, “Sequencing and optimization within an embodied task dynamic model,” *Cognitive Science*, vol. 35, no. 3, pp. 527–562, 2011.

[3] A. Turk and S. Shattuck-Hufnagel, *Speech timing: Implications for theories of phonology, speech production, and speech motor control*. Oxford University Press, USA, 2020.

[4] ——, “Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production,” *Frontiers in Psychology*, p. 2952, 2020.

[5] D. N. Lee, “Guiding movement by coupling taus,” *Ecological psychology*, vol. 10, no. 3-4, pp. 221–250, 1998.

[6] C. P. Browman and L. Goldstein, “Articulatory phonology: An overview,” *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[7] E. Todorov and M. I. Jordan, “Optimal feedback control as a theory of motor coordination,” *Nature neuroscience*, vol. 5, no. 11, pp. 1226–1235, 2002.

[8] K. D. Roon and A. I. Gafos, “A dynamical model of the speech perception-production link,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35, no. 35, 2013.

[9] D. Byrd and E. Saltzman, “The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening,” *Journal of Phonetics*, vol. 31, no. 2, pp. 149–180, 2003.

[10] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, “Model-based reproduction of articulatory trajectories for consonant–vowel sequences,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.

[11] T. Sorensen and A. Gafos, “The gesture as an autonomous nonlinear dynamical system,” *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, 2016.

[12] B. Elie, D. Lee, and A. Turk, “Modeling trajectories of human speech articulators using general tau theory,” *Speech Communication*, to appear.

[13] J. S. Perkell, “Movement goals and feedback and feedforward control mechanisms in speech production,” *Journal of neurolinguistics*, vol. 25, no. 5, pp. 382–407, 2012.

[14] F. H. Guenther, “A neural network model of speech acquisition and motor equivalent speech production,” *Biological cybernetics*, vol. 72, no. 1, pp. 43–53, 1994.

[15] H. Nam, L. Goldstein, and E. Saltzman, “Self-organization of syllable structure: A coupled oscillator model,” *Approaches to phonological complexity*, vol. 16, pp. 299–328, 2009.

[16] S. Tilsen, “A dynamical model of hierarchical selection and coordination in speech planning,” *PloS one*, vol. 8, no. 4, p. e62800, 2013.

[17] ——, “Three mechanisms for modeling articulation: selection, coordination, and intention,” *Working Papers of the Cornell Phonetics Laboratory*, 2018.

[18] J. Simko, “The embodied modelling of gestural sequencing in speech,” University College Dublin. School of Computer Science and Informatics, Tech. Rep., 2009.

[19] E. Saltzman, “Task dynamic coordination of the speech articulators: A preliminary model,” *Experimental brain research series*, vol. 15, pp. 129–144, 1986.

[20] E. L. Saltzman and K. G. Munhall, “A dynamical approach to gestural patterning in speech production,” *Ecological psychology*, vol. 1, no. 4, pp. 333–382, 1989.