A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing

Ralf Winkler, Liang Ma, Pascal Perrier

A model of optimal speech production planning

integrating dynamical constraints to achieve appropriate articulatory timing.

Ralf Winkler[1,2,3], Liang Ma[1,4], Pascal Perrier[1]

[1]DPC/GIPSA-lab, Grenoble, France
[2]ZAS/Phonetik, Berlin, Germany
[3]Technische Universität, Berlin, Germany
[4]Zhejiang University, China

INTRODUCTION

It has been often suggested that, similarly to skilled arm or limb movements in humans, the production of speech gestures could be based on an optimal planning in the central nervous system. This planning would use internal representations [1,2] of the speech production apparatus [3,4] to determine the motor command patterns allowing the achievement by the speaker of the desired speech communication goals with the minimum of effort. In this context major issues are related to the nature of the speech communication goals (targets or spatiotemporal trajectories) and to the definition of minimum of effort (minimum motor command change, minimum velocity peak, trajectory smoothness, minimum of jerk...). GEPPETO[1], the speech production model presented in this paper, has been designed within this general theoretical framework [5]. It is based on a motor control model involving optimal planning to shape a biomechanical model of the vocal tract coupled with a harmonic acoustical model of speech production. It will be shown how dynamical constraints can be taken into account in order to achieve appropriate articulatory timing.

THE GEPPETO MODEL

In GEPPETO, speech goals are linked with a segmental description of the phonological input in terms of phonemes. Thus, goals are related to phonemes and they are specified as 3D ellipsoids in the acoustic space of the first three spectral maxima (F1, F2, F3) of the vocal tract transfer function. These ellipsoids are determined by their centres, considered to be canonical acoustic realisations of the phonemes, and by standard deviations along the three directions F1, F2, F3. These standard deviations are assumed to account for the acoustic variability that is tolerated around each canonical realisation without any consequence on the auditory perception. The motor control model includes a muscle force generation mechanism based on the Equilibrium-Point hypothesis (Feldman, 1986). Thus, for a given muscle, the motor control variable is the muscle length threshold above which active muscle force is produced. Movements are generated by shifting the motor control variables at a constant rate of shift between motor targets. In GEPPETO, the motor targets are associated with the speech goals in the acoustic domain. These motor commands are sent to the seven muscles of a 2D biomechanical model of the tongue [6] that is embedded in a 2D description of the vocal tract boundaries. The tongue model deforms and moves as the result of the combination of the influence of the target motor commands, of their timing and of the dynamical properties of the model (muscle forces, tissues elasticity, friction and contacts with external structures). For a given speech sequence, the target motor commands associated with each of the phonemes in the sequence are selected thanks to an optimal planning in which the minimum of effort is considered to correspond to the closest possible neighbourhood between all the targets commands of the sequence. Hence, this optimal planning does not rely on any characteristics of the articulatory movement between two targets and is not related to articulatory trajectories. Here, optimal planning is achieved by minimization of distances in the motor control variable space (speaker-related requirements) subjected to listener-related constraints (ellipsoids in the space of the three first formants).

In a first version of the model [5,7], optimization with listener-related constraints ensured to select motor commands that are associated with spectral patterns located inside the phoneme related target-ellipsoids. Taking the perceptual constraint into consideration was possible by the use of a static forward internal model that associates motor control variables and F1-F2-F3 spectral patterns. The

[1] GEPPETO holds for "**GE**stures shaped by the **P**hysics and by a **PE**rceptually oriented **T**argets **O**ptimization"

static internal model was trained beforehand. This approach has been shown to be efficient to account for coarticulation phenomena and in particular for anticipatory behaviours. However, it did not offer any way to ensure that the intended target spectral patterns are actually reached for a certain timing of the command. Indeed, it is known that articulators move more or less fast depending on their dynamic characteristics. Thus, a stronger and more realistic listener-related constraint for the optimal planning is the selection of motor control variables compatible with the actual achievement of the target spectral patterns for a given timing of the command. It is proposed in this study that this constraint corresponds to a dynamical constraint expressed in terms of global force level. The value of global force level was defined as the sum of forces values generated the different muscles..

More specifically, in the GEPPETO model, optimal planning is now constrained by the double necessity to select motor commands appropriate to the achievement of target spectral patterns inside the target ellipsoids and to ensure that the global level of force remains within a given range during the whole movement. The choice of this range is guided by the timing of the commands (i.e. the speaking rate) and by perceptual accuracy requirements: for slow speaking rates or low accuracy requirements, a low level of force can be used; for fast speaking rates and great accuracy a strong level of force is required. To include this dynamical constraint in the optimal planning process, a second internal forward model has been learned. It associates the motor commands with the corresponding global muscle force level. We call it "*dynamical forward model*". Thus, the enhanced optimal planning presented in this paper minimizes the size of the neighbourhood defined by successive target commands while respecting both listener-related constraints. In the current model, optimization is accomplished by sequential quadratic programming (SQP), which allows to closely mimic Newton's method (Quasi-Newton) for constrained optimization. In the current state of the model three ranges of global force level have been defined: low, normal and high. The corresponding force values were derived from a database of 8293 tongue movements simulated with the 2D biomechanical model from the rest position to a randomly chosen target. The total force value was computed for all targets. Then, the minimum and maximum forces were extracted together with an intermediate value, and they were used as levels for the three force ranges. They were used as starting points in the optimization procedure. As mentioned above, the newly incorporated dynamical constraint allows the optimization algorithm finding the best set of motor commands for the production of the sequence while remaining within the range of predefined total force. During optimization the mean total force value was considered to be within the predetermined force range if its rise or its fall did not excite 0.5 Newton

RESULTS and DISCUSSION

The new optimal planning process has been tested so far on VCV sequences, starting from the tongue rest position. Optimal motor command patterns were found for each segment of the sequences for each of the three force constraints. Then, the 2D biomechanical tongue model was used to simulate tongue movements for these 3 force levels and for three timing of the motor commands: slow, normal and fast. Simulations are currently in progress. The first results show that the global muscle force can have a significant impact on the articulatory trajectories, in terms of curvature and in terms of positions actually reached at targets. At a slow speaking rate, changing the global level of force tends to have less consequence on the articulatory positions reached at targets than at high speaking rates, but it seems to have an impact on the trajectory shape.

The preliminary results support the original hypothesis that applying the appropriate dynamical constraints in the optimal planning process helps dealing with articulatory timing and perceptual accuracy expressed in terms of articulatory positions at targets. In this perspective, contrary to the extrinsic timing theories defended among others by Fowler [8] or Kelso et al. [9] time control could be seen as a combination of centrally specified and physically constrained characteristics

REFERENCES

[1] Jordan, M.I. (1990). Motor Learning and the Degrees of Freedom Problem. In M. Jeannerod (ed.), *Attention and Performance*, Hillsdale, NJ: Erlbaum, pp. 796-836.

[2] Kawato, M., Maeda, Y., Uno, Y. & Suzuki, R. (1990) Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. *Biological Cybernetics, 62,* pp. 275-288.

[3] Guenther, F.H., Hampson, M. & Johnson, D. (1998) A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105,* pp. 611–633.

[4] Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech Communication 22*(2-3), 251-267.

[5] Perrier, P. & Ma., L. (2008). Speech planning for V_1CV_2 sequences: Influence of the planned sequence. *Proceedings of the 8th International Seminar on Speech Production (ISSP 2008)* (pp. 69-72). Université de Strasbourg, France.

[6] Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. (2003) Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *J. Acoust. Soc. Am., 114(3),* pp. 1582-1599.

[7] Perrier, P., Ma, L. & Payan, Y. (2005) Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue*. Proceedings of Interspeech 2005*, Lisbon, Portugal, pp. 1041-1044.

[8] Fowler, C.A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics 8*, 113-133.

[9] Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E. & Kay, B. (1985). A qualitative dynamical analysis of reiterant speech production: Phase portraits, kinematics and dynamic modeling. *Journal of the Acoustical Society of America*, 77, 266-290.