RELATIVE TIMING OF BILABIAL GESTURE IN FINNISH

Michael L. O'Dell[a], Juraj Šimko[b], Tommi Nieminen[c], Martti Vainio[d] & Mona Lehtinen[d]

[a]University of Tampere, Finland; [b]Bielefeld University, Germany;
[c]University of Turku, Finland; [d]Univeristy of Helsinki, Finland
michael.odell@uta.fi; juraj.simko@uni-bielefeld.de; tommi.nieminen@utu.fi;
martti.vainio@helsinki.fi; mona.lehtinen@helsinki.fi

ABSTRACT

The Embodied Task Dynamic model of gestural sequencing predicts that an intervocalic consonantal lip closing gesture should come later if the tongue is moving from /i/ to /a/ rather than from /a/ to /i/ because this relation is more efficient in terms of production and perceptibility. We tested this prediction for Finnish /ipa/ and /api/ using EMA to track articulation. The results confirmed the predictions of the model for single /p/ and also revealed a significantly greater lag for geminate /pp/. This quantity effect is also born out by the model.

Keywords: gestural timing, embodied task dynamics, Finnish, quantity, EMA

1. BACKGROUND

1.1. Articulatory phonology and gestural score

In Articulatory Phonology (AP), an utterance is fully described by its gestural score [1]. Active gestures drive vocal tract articulators towards target positions corresponding to so called tract variables that represent the vocal tract state relevant to the given task. The lip aperture (LA) tract variable, for example, captures the distance between the lips and is thus linked to the degree to which the task of a bilabial closure is achieved at a given moment.

The dynamics of the vocal tract under the influence of an active gesture is usually modeled using the Task Dynamics (TD) theory of target-oriented motor action. The behavior of each tract variable involved in achieving the given gestural target is obtained as a solution of a damped mass-spring dynamical system which has the given target as its equilibrium position and a stiffness parameter determining the responsiveness of the system to the given task [9].

The temporal details of gesture activation onsets and offsets, plus the gestural stiffness parameters, are the sole factors governing the surface form of a gestural sequence—the intended utterance. The question of how these parameters are determined must thus be central to any inquiry into the nature of speech production. Browman and Goldstein [2] proposed a rule-based account of intergestural timing. The relative onsets and offsets of the neighboring gestures depended solely on the type of their mutual lexical affiliation. In the subsequent work [3, 10], a coupled oscillator methodology has been used to refine this local approach and generalize it to global sequencing patterns.

1.2. Embodied task dynamics

Šimko and Cummins [12] proposed an alternative account of gestural timing called Embodied Task Dynamics. Inspired by Lindblom's Hypo-hyperarticulation (H&H) and Emergent Phonology theories [5, 6], their model is based on the idea that local and global intergestural relations depend on optimality principles. That is, the timing details of a gestural sequence as well as the stiffness parameter values of participating gestures are optimal with respect to competing production and perception efficiency requirements.

1.2.1. Cost and efficiency

Competing efficiency requirements are represented by three cost functions: articulatory effort *E*, parsing cost *P* and duration cost *D*. An overall cost function whose minima are presumed to represent optimal gestural scores is then defined as a weighted sum of these three components:

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

where the weight coefficients α_E, α_P and α_D represent high level intentional parameters that can impose sequencing modifications corresponding to lax and tense articulation and speaking rate.

Articulatory effort *E* is calculated as an integral of forces exerted by model muscles in order to move the vocal tract articulators through the given sequence of gestural targets. In order to approxi- mate these forces, the TD model has been extended by linking the target oriented behavior of the speech system with the underlying physical and anatomical constraints. The articulators thus act as mass springs *with realistic mass parameters*, in contrast to the traditional TD implementation where the mass of articulators is nominal (unit mass). (For implementation details, see [11]).

Parsing cost *P* is an approximate measure of the listener's effort in perceiving an utterance. Šimko and Cummins [12] presume that this effort is related to the articulatory precision of each realized gesture in the utterance. Furthermore, gestures with longer realization intervals are presumed to be perceived more easily by the listener than shorter ones. This is modeled using a monotonically increasing function of time called the duration estimate function $d_e(t)$ starting at 0 and asymptotically converging to 1 during the realization interval. The overall parsing cost associated with the whole gestural sequence is thus calculated as a function of both the precisions and the durations of individual realized gestures. It increases as precision decreases ("undershoot") and decreases for longer durations. As was the case with the cost *E* expressing the articulatory effort, the value of *P* is again a function of the onset and offset times of gestural activation intervals, and of the value of the overall stiffness parameter.

Duration cost *D*, the third cost component, simply evaluates the overall duration of gestural sequence realization. It is computed as the length of the interval from the onset of the first gesture in the sequence to the offset of the realization interval of the last gesture (in seconds). While the parsing cost *P* captures the position of the given utterance on the H&H scale, the value *D* reflects speaking rate.

1.2.2. Finding an optimal gestural score

To identify the gestural score optimal with respect to the combination of these components, an optimization procedure based on simulated annealing can be used [11]. The function *C* (see above) mapping the onsets and offsets of gestural activation intervals and the overall stiffness value to the quantitative cost measure is the objective function of this optimization problem. The weight coefficients αE, αP and αD are fixed parameters of the objective function representing the intentions of the speaker with respect to the H&H scale and speaking rate. All other parameters are seen as speaker or language dependent. Note that the optimal gestural score/stiffness combination emerges as the result of the cost optimization alone. Only the sequence of "perceived" gestures is imposed; no explicit phonological rules governing the relative timing of gestures are used.

1.3. A specific prediction

The Embodied Task Dynamic model makes many specific predictions. In this study we focus on one non-trivial local phasing prediction. One robust result of the optimization simulations that have been carried out is that tongue body movement (vowel transition) should start later for /a/ → /i/ than for /i/ → /a/ in relation to an intervocalic bilabial stop gesture, e.g. in /api/ tongue body movement (vowel transition) should start later relative to the lip clos- ing gesture than it does in /ipa/. Just such a relation was in fact observed by Löfqvist & Gracco [8] for English speakers. Embodied Task Dynamics offers an explanation for this state of affairs in terms of efficiency: simulations show that this phasing relation is optimal in terms of using minimal energy for production while maintaining perceptibility [11, 12].

In the present study we had two specific research questions: (1) *Does this gestural phasing relation hold true for Finnish?* and (2) *Does quantity have an effect on tongue body-to-lip phasing?*

2. METHODS

2.1. Articulography (EMA)

In order to address the question at hand, we used electromagnetic articulography (EMA) to track articulatory movements of native Finnish speakers. The apparatus used for this purpose was the Carstens AG 500 Articulograph at the University of Helsinki. Simultaneous audio recordings were made while the articulograph recorded the three-dimensional movements of sensors attached to subjects' articulators.

Sensors were monitored attached to the upper lip, the lower lip, the tongue body, the tongue tip, as well as three reference points: behind each ear and at the bridge of the nose.

2.2. Test material

Subjects first pronounced the test words *tati*, *tapi*, *tita*, *tipa*, *tatti*, *tappi*, *titta*, *tippa*, *tipta*, *tapti* embedded in the carrier sentence *Mitä sana* ___ *tarkoittaa?* ("What does the word ___ mean?"). Subjects read five blocks of these sentences in pseudorandom order, took a pause, then read five more blocks, then after another pause read five more blocks. Thus each sentence was read a total of 15 times. We report here only on the words with intervocalic /p/ or /pp/ (*tapi, tipa, tappi, tippa*).

After this subjects pronounced sequences of four test words (*tipa tipa ...*, *tapi tapi ...*, *pati pati ...*, *pita pita ...*) continuously for approximately ten seconds each, resulting in approximately 30 cyclic repetitions for each word.

In the following we report results for one female speaker (preliminary results for a second speaker are very similar).

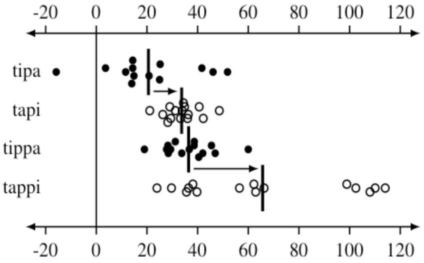
2.3. Measurements

Following Löfqvist & Gracco [8] we made two measurements each for the test words in carrier sentences. A lip aperture curve was computed as the vertical distance between lip sensors and the onset of the lip closing gesture was taken defined as the point of zero aperture velocity prior to lip closure. Likewise a tongue body speed curve was calculated from the horizontal and vertical components of the tongue body sensor and onset of the tongue body gesture was defined as the point of minimum speed prior to the vowel transition.

3. RESULTS

3.1. Sentences

Figure 1: Tongue body lag (ms) for sentences.

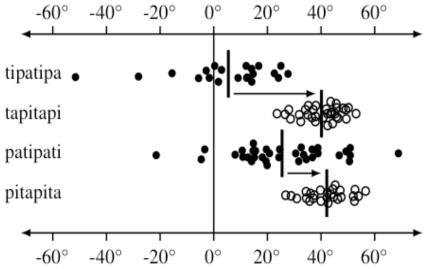


Results for the test words in carrier sentences are shown in Fig. 1. Each dot indicates the measured tongue body lag (tongue body onset time minus lip onset time) for a single test word token. The vertical lines indicate the mean lags. As predicted, tongue lag is greater on average for tapi compared to tipa as well as tappi compared to tippa. It also appears that the geminate stops in both cases have a greater lag on average compared to the corresponding singleton stop. These differences are significant statistically, although there is considerable overlap in the distributions (ANOVA: vowel effect F(1, 54) = 17.22, p < 0.001, quantity effect F(1, 54) = 20.56, p < 0.001, interaction F(1, 54) = 2.43, n.s. Identical results were obtained using Bayesian inference for means, in addition variance was greater for tipa than for tapi and less for tippa than for tappi, p < 0.001).

3.2. Cyclic repetitions

Results for the cyclic repetitions are shown in Fig. 2. Dots indicate tongue body onset relative to lip onset in degrees (360 ° = one cycle) with mean angle indicated by vertical lines. The cyclic repetitions gave results parallel to the sentences, that is, tongue lag is greater on average in the *api* contexts than in the corresponding *ipa* contexts. In addition it is obvious that there is much less variation in phase difference for the *api* cases (roughly 40 °), indicating that the two gestures are more tightly coupled. Differences in means and variances were both significant (p < 0.001 using Bayesian inference).

Figure 2: Tongue body lag (degrees) for cyclic repetitions.

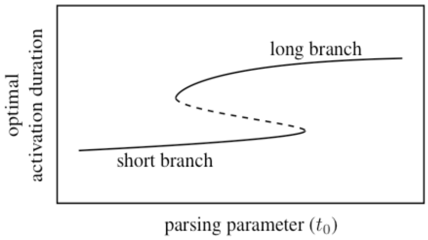


We speculate that in the *api* case, where the lower lip must rise for lip closure at approximately the same time as the tongue is rising from /a/ to /i/, there is a considerable gain in efficiency if the jaw is raised as well. In the *ipa* case, where lower lip and tongue body need to move in opposite directions, possible gains in efficiency should be smaller, since any jaw movement will aid one articulator but work against the other.

3.3. Effect of quantity

For our Finnish speaker tongue lag was greater for geminates than for singletons. To answer the question of whether or not this effect could also be explained in terms of efficiency we ran additional optimizations using the Embodied Task Dynamic model, systematically varying the duration estimate function of the parsing cost (see above) using an additional parameter t_0 to control how quickly the function increases: $d_e(t/t_0)$. If the duration estimate function is interpreted as the probability that a listener perceives the gesture in question, t_0 corresponds to the duration at which probability of perception is 50 %. As t_0 increases, biasing the system towards longer durations, the optimal gesture activation for lip aperture jumps to a longer value. At intermediate values of t_0 there is a region of bistability where two locally optimal activation durations coexist (though the shorter one is globally optimal). This is to say there appears to be a natural bifurcation of activation durations with hysteresis (pair of saddle-node bifurcations, see Fig. 3).

Figure 3: Schematic hysteresis. Solid line indicates local minimum of overall cost function.



Furthermore, the lip to tongue lag is indeed greater for the longer branch, mimicking the empirical situation shown in Fig. 1. This relation holds for the /i...a/ context (*tipa* vs. *tippa*) as well as the /a...i/ context (*tapi* vs. *tappi*).

We also looked at the kinematic measures used by L öfqvist [7] to investigate geminates. In general our empirical results are in agreement with Löfqvist's results for Japanese, as well as model simulations for short (C) vs. long (CC) realizations (see Table 1). Especially notable is that, in accordance with the data, the model predicts greater lip compression for geminates (even though identical /p/ targets are assumed).

Table 1: Relation of C to CC for Japanese (= *J*, L öfqvist [7]), Finnish (= *F*, present study) and model optimization (= *m*). [< (>): C less (greater) than CC with significance */**/***: p < 0.05/0.01/0.001]

measure	*J*	*F*	*m*
closure duration	<***	<***	<
u-lip velocity	>**	n.s.	≈,>
l-lip position	<***	<**	<
l-lip displacement	<***	<*	<
l-lip velocity	>n.s.	n.s.	≈
l-lip stiffness	>***	>***	>

One aspect of the model optimization is contrary to our Finnish data: The model systematically selects earlier lip activation for geminates, resulting in shorter realization of the preceding vowel. This effect is more pronounced for the /a...i/ context. Finnish on the other hand typically shows a small lengthening of the vowel preceding geminates (in the present data the vowel is significantly longer before geminate /pp/ in both vowel contexts). Interestingly this inverse relation for consonant and preceding vowel does hold for some other quantity languages such as Italian [4]. This difference is likely related to organizational differences not incorporated in the Embodied Task Dynamics model, similar to those discussed in [13] for Italian and Japanese.

4. SUMMARY

Our research has confirmed one prediction of Embodied Task Dynamics, at least for one Finnish speaker (preliminary examination of a second speaker indicates similar results): Tongue body movement (vowel transition) does start later on average for /a/ → /i/ than for /i/ → /a/ in relation to lip onset for intervocalic /p/. Besides providing further support for the model, this is interesting in and of itself as an extention to Finnish of the phenomenon reported by L öfqvist and Gracco for English [8]. In addition the data indicated a larger tongue body lag for geminate /pp/, which the Embodied Task Dynamics model also predicts, assuming quantity is dependent on duration perception.

5. REFERENCES

[1] Browman, C.P., Goldstein, L. 1989. Articulatory gestures as phonological units. *Phonology*, 6, 151-206.

[2] Browman, C.P., Goldstein, L. 1990. Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J., Beckman, M.E. (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* Cambridge: Cambridge University Press, 341-376.

[3] Browman, C.P., Goldstein, L. 2000. Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parl ée* 5, 25-34.

[4] Farnetani, E., Kori, S. 1986. Effects of syllable and word structure on segmental durations in Italian. *Speech Communication* 5(1), 17-34.

[5] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H & H theory. In Hardcastle, W.J., Marchal, A. (eds.), *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 403-439.

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

[7] L öfqvist, A. 2005. Lip kinematics in long and short stop and fricative consonants. *JASA* 117(2), 858-878.

[8] L öfqvist, A., Gracco, V.L. 1999. Interarticulator program- ming in VCV sequences: Lip and tongue movements. *JASA* 105(3), 1864-1876.

[9] Saltzman, E.L., Munhall, K.G. 1989. A dynamical approach to gestural patterning in speech production. *Ecological* 1(4), 333-382.

[10] Saltzman, E.L., Nam, H., Krivokapić, J., Goldstein L. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In Barbosa, P.A., Madureira, S., Reis, C. (eds.), *Speech Prosody 2008: Fourth Conference on Speech Prosody* Campinas, Brazil, 175-184.

[11] Šimko, J. 2009. *The Embodied Modelling of Gestural Sequencing in Speech*. University College Dublin, Tech. Rep.

[12] Šimko, J., Cummins, F. 2010. Embodied Task Dynamics. *Psychological Review* 117(4), 1229-1246.

[13] Smith, C.L. 1995. Prosodic patterns in the coordination of vowel and consonant gestures. In Connell, B., Arvaniti, A. (eds.), *Papers in Laboratory Phonology IV, Phonology and Phonetic Evidence* CUP, 205-222.