The Supervision of Speech Production

Mark Tatham

Reproduced from Schoentgen, J., Ramlot, J-M, Sorin, C., Méloni, H. and Mariani, J. (eds.), Levels in Speech Communication: Relations and Interactions (1995). Amsterdam: Elsevier, pp. 115-125.

Introduction

Since the emergence of phonology as a component within the generative paradigm of linguistics there has been discussion concerning how it relates to phonetics. At least two views of the relationship are possible:

- Phonology describes the same data as phonetics, but whereas phonetics models physical detail phonology models abstract relationships holding between units within the data. These units are defined within the general framework of linguistic theory.
- Phonology stands logically prior to phonetics, and the output of its processes comprises the input to phonetics. Phonetics itself is a processing component producing an output. The model can be regarded as static, involving no temporal relationship (the usual view in linguistics), or it can be regarded as dynamic. The dynamic view is often taken when the theory is used to support work in an allied area, for example in speech technology, where temporal as well as logical relationships between processes are important.

It is usually understood that phonetic processes are of little concern to linguistics, and that by the output of phonology all linguistic processing is complete. Phonetic realisation of phonological 'requirements' is thought of as a passive process involving no cognitive processing, and introducing nothing new of linguistic consequence or interest. For example, phonologists working on language acquisition are interested in the phonetic constraints on what can be acquired by a child, but are not concerned with the detail of phonetic processes. The most extreme form of this position would be that speech processing of a cognitive nature falls within the province of phonology, but that all physical processing falls within the province of phonetics. Since, by definition, language is a cognitive system it can have nothing formal to do with phonetics, except in its trivial realisational rôle.

There has recently been a trend toward a more phonetic approach to phonology,[1,2] and a greater awareness among phoneticians of the abstractions ofphonology3 and why they are necessary. This paper will consider how ARTICULATORY PHONOLOGY might handle some phonetic data, and attempt to deal with some problems that the data points up in the theory.

Articulatory Phonology

Described first by Browman and Goldstein,[2] ARTICULATORY PHONOLOGY aims to unify phonetics and phonology by treating them as 'low and high dimensional descriptions of a single system'. The aim is partly achieved in the claim that the constraints of the physical system underlie the phonological system, and partly by making the units of control at the planning level the same as those at the physical level. The idea is to blur the distinctions between phonology and phonetics, and between the planning and execution of an utterance. The phonology provides an input to a TASK DYNAMIC MODEL of speech production.6 The utterance plan input to the task dynamics takes the form of a *gestural score* (examples of which are shown later in Fig. 3).

The gestural score sets out the locations and degrees of constrictions within the vocal tract and their timing during the progression of the utterance. Sequencing and durations of gestures are critical to the score as are the temporal relationships between the various *tract variables* involved. The tract variables are descriptive parameters of the vocal tract which can be manipulated in the TASK DYNAMIC MODEL by the involvement of articulator groupings.

The variables include, for example, lip aperture, tongue tip constriction degree, tongue body constriction location, velar aperture, glottal aperture, etc. So, the score for the single-sound utterance [æ] might show that for a certain time the tongue body constriction is to be in the area of the pharynx and wide, with the velar aperture closed to prevent nasality, and the glottis closed* to allow vocal cord vibration; other tract variables mayor may not be specified. The gestural score, however, remains *a plan.* That is, it is abstract and is not intended to be a description of the actual vocal tract movements. For this reason score gestures are discrete, not continuous.

* Throughout this paper the term *closed glottis* refers to a state of adjustment of vocal cord tension that is appropriate for spontaneous vibration, given the current sub glottal air pressure. The term *open glottis* refers to lack of vocal cord tension.

In the TASK DYNAMIC MODEL gestures have a functional goal (the task) which is achieved by *coordinative structures,*[7] internally communicating groupings of articulators or their underlying musculature. It is primarily the expression of functionality which characterises the model's phonological perspective, and the task specification which characterises its phonetic perspective. Each of the tasks is independent (though related functionally *via* the gestural score), the dynamic aspect of the model being the control of movement towards particular physical goals. The model focuses on the task itself rather than on parts of the articulatory system involved in executing the task.

Some Phonetic Data – I

Fig. 1 shows data from an experiment which simultaneously recorded intraoral air pressure measurements and electromyographic (EMG) signals from the lip musculature during the production of bilabial stops.[8] The data is from examples of *a purr* and *a burr.* Tracings from top to bottom are

- intraoral air pressure (low pass filtered at 50Hz),
- smoothed (25ms effective integration time) surface electrode EMG from *m. orbicularis oris* (responsible for lip closure), *m. quadratus labii (superiori* and *inferiori)* and *m. mentalis* combined (responsible for lip opening),
- the associated acoustic signal (to 5kHz).

For illustrative purposes the examples were chosen as having nearly identical closure times for the stops, but in all respects they are very typical of the 25 examples of each utterance provided by the speaker of southern British English. Vertical lines on each example show the moments of closure and opening of the lips for the stops.

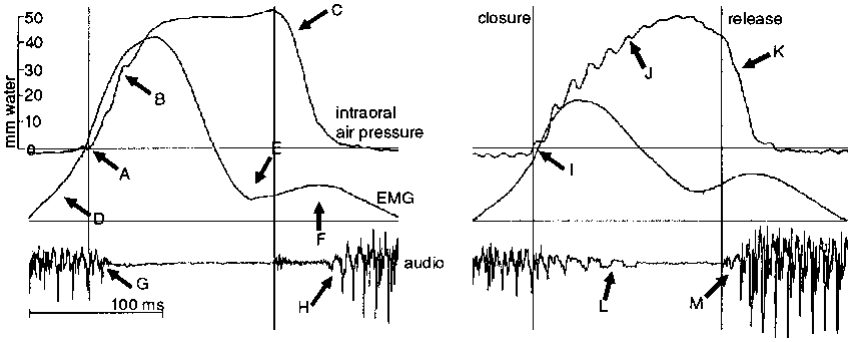

Fig. 1 a *purr* and a *burr*- air pressure, emg and audio.

The following points should be noted from the example of *a purr*

- intraoral air pressure shows a sharp rise at the moment of closure - A

- some vocal cord vibration is visible in the pressure curve after closure - B
- following stop release pressure falls sharply for about 30ms before levelling off to reach zero by about 50ms. Levelling off correlates with onset of vocal cord vibration - C
- there is EMG activity (from m. orbicularis oris) for at least 50ms before lip closure - D
- EMG activity (from m. orbicularis oris) reaches its post-closure Iowa few milliseconds before closure release, as the other muscles increase activity to achieve lip opening - E
- EMG activity from the opening muscles continues about 80ms after release - F
- two cycles of vocal cord vibration are visible in the waveform after lip closure - G
- vocal cord vibration associated with the vowel is visible in the waveform about 45ms after the release of [p] - H.

And the following points should be noted from the example of *a burr*

- intraoral pressure rises with closure, though the angle of rise is less than in a purr - I
- vocal cord vibration is visible during the first three-quarters of the closure period - J
- pressure falls sharply for about 25ms following release, reaching zero shortly after - K
- except for its peak amplitude (and its timing with respect to moment of closure) the EMG signal is similar to that associated with a purr; for the 25 samples of data there was no significant difference in peak amplitude of the EMG signal between a purr and a burr
- five cycles of vocal cord vibration are visible in the waveform after lip closure - L
- vibration following the release is visible in the audio after less than 10ms - M.

The differences in amplitude and timing of the peak of the EMG signal are explained by the natural variability often observed in EMG signals.[9] This 1973 study concluded that from the point of view of muscular tension in the lips (the primary articulator associated with bilabial stops) there was no significant difference between [p] and [b].

Two intraoral air pressure events are of interest here:

- in *a burr,* the modulation of the signal throughout with vocal cord vibration, and
- the lower peak amplitude reached during the [b] closure - peak amplitudes for [b] and [p] in the entire data set being in a ratio of 1:1.25 .

In [b] vocal cord tension and the pressure drop across the glottis are such that vibration has been sustained during most of the closure - even though the audio shows no more than half a dozen cycles. Because of this glottal vibration a simple binary description of the glottis here would be *closed* (rather than *open).* Equally clearly the glottis would be described as open during the closure for [p] despite the overhang of vibration associated with the preceding vowel.

The acoustic signal shows the characteristic delay in vocal cord vibration for the vowel following [p] in English. This is usually explained by the fact that although the glottis is described as closed during this period the pressure differential below and above the vocal cords, for a given vocal cord tension, is insufficient for vibration to begin. It does so when the supra-glottal pressure has fallen to the correct level. The co articulatory effect is due also to the rapidly falling pressure and the fact that the fall often terminates in an oscillatory motion involving cycles of negative pressure (not visible in this example). Glottal vibration must wait not just for the right supra-glottal pressure but for the aerodynamic system to regain stability following these oscillations.

But why was there not such a delay following [b]? The answer is that the pressure differential and delay before normalisation were *different* from those associated with [p]. The

clue to the explanation lies in the shape of the intraoral air pressure curve during the closure: the peak is reached earlier than with [p], and there is a significant pressure downturn before release. The claim here is that in addition to the fact that during [b] the glottis is closed the speaker is doing something special to reduce the intraoral air pressure

- to keep the vocal cord vibration going as long as possible during closure, and
- to keep it going into the vowel, or at least to get it restarted as soon as possible after the release of the stop.

In other words the aerodynamic coarticulatory effect observed in association with [p] is here being *constrained* – the DYNAMIC SPEECH SCENARIO is being *supervised*. We shall return to these concepts after the next set of data.

Some Phonetic Data – II

Fig. 2 shows waveforms (to 5kHz) of the French utterances *une panne* and *une banne*, and the English utterances *a pan* and *a ban*. The examples illustrated are from a large data set and have been selected because their bilabial stops have similar closure durations. This is simply for illustrative purposes to enable alignment of closure and release of the stops. They are nevertheless typical examples from a speaker of 'standard' French and a speaker of 'standard' British English, and the observations below are true for examples with differing closure durations. This is contemporary data.

We can compare what is happening to vocal cord vibration surrounding the stop consonants in all four utterances. Of course, there are differences between the two languages which constrain the comparison - the vowels in the nouns have different articulatory and acoustic properties, the 'pronounced' *e-muet* of *une* is not the same as the English word *a*, and the rhythmic systems of the two languages are different. Despite these facts I believe the comparisons to be made are valid.

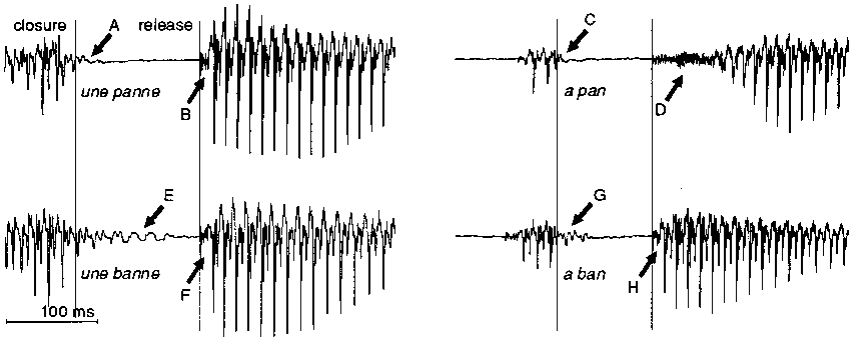

Fig. 2 une panne, une banne, a pan, a ban – audio.

Notice in the acoustic signal for une panne

- there is some overhang of vocal cord vibration into the closure phase of the stop
- vocal cord vibration associated with the vowel begins shortly (around 10ms) after the stop release; note, though, that the vibration itself has around a 10ms period.

And in the acoustic signal for *a pan*

- there is perhaps one cycle of vocal cord vibration overhang into the closure
- vibration associated with the vowel is delayed by around 40-50ms following the release.

Notice in the acoustic signal for *une banne*

- audible vocal cord vibration continues at least three quarters of the way through the closure phase of the stop
- following release, one vocal cord cycle may be lost or interrupted.

And in the acoustic signal for *a ban*

- vocal cord vibration extends about two cycles into the closure
- following the stop release there is around 10ms delay in vocal cord vibration onset, though the period of the following vibration is itself about 10ms.

We note then that $[p]_{Eng}$ is different from $[p]_{Fr}$ and that $[b]_{Eng}$ is different from $[b]_{Fr}$. But we also note that as far as the data presented here is concerned there are enough similarities between $[b]_{Eng}$ and $[p]_{Fr}$ to set up a cross-language descriptive acoustic phonetic model incorporating three rather than four objects. These are: $[p]_{Eng}$, $[b]_{Eng}$/$[p]_{Fr}$, and $[b]_{Fr}$. English uses the first two to realise the phonological contrast between /p/ and/b/, whereas French uses the second two for its /p/-/b/ contrast. We might say that these acoustic realisations are part of the set of goals of speech production in the two languages.

But to do so conceals at what stage in speech production processing the functional identity of $/p/_{Eng}$ and $/p/_{Fr}$, and $/b/_{Eng}$ and $/b/_{Fr}$ switches from a two-object system to a three-object system. The previous data involving intraoral air pressure measurements pointed toward glottal closure during [b] in English, and if this is so, then it is certainly the case (even from the acoustic data) that there is glottal closure during [b] in French. Both sets of data point toward glottal opening in [p] for English, and the acoustic data points toward glottal opening in [p] in both English and French.

In terms then of glottal gesture we have $[p]_{Eng}$ and $[p]_{Fr}$ sharing the same goal of opening, and $[b]_{Eng}$ and $[b]_{Fr}$ the goal of closure. This corresponds to the more abstract phonological arrangement in the two languages – two objects, not three. So, the system adopted in the phonology continues at least as far as physical glottal gesture, taking in motor control along the way because opening and closure would not occur without the appropriate motor control of the coordinative structures involved.

What remains then is to account for the different behaviour of the two languages at the acoustic level compared with their behaviour at the articulatory gestural level. At this point we turn to the theory of coarticulation. The delay in vocal cord vibration onset following the release of $[p]_{Eng}$ is a well documented aerodynamic effect (described above). Based on this explanation we also expect a similar delay to occur following $[b]_{Eng}$. But this is *not* what happens. Neither does it happen for $[p]_{Fr}$ or $[b]_{Fr}$.

The vibration damping during the closure phase of $[b]_{Eng}$ is also the result of the aerodynamics of the system: as the supraglottal air pressure rises towards the value of the subglottal air pressure, vocal cord vibration will tend to decrease in amplitude and then stop altogether. Similarly we would expect early failure of vibration during the closure $[b]_{Fr}$. But this is *not* what happens: it carries on, often throughout the closure.

So $[p]_{Eng}$ is not the same as $[p]_{Fr}$, and $[b]_{Eng}$ is not the same as $[b]_{Fr}$ except at the superficial acoustic level. If they were the same the aerodynamic constraints placed on both would produce similar acoustic results. But having concluded that there must be a difference it is clear that it does not reside in the planned glottal gestures – those seem to be the same for both languages.

So, we leave our data with the following observations for /p/, [p], /b/ and [b] in word-initial context in English and French:

p_{Eng} - p_{Fr}

- phonology - identical
- gestural plan - identical
- acoustic signal - different

b_{Eng} - b_{Fr}

- phonology - identical
- gestural plan - identical
- acoustic signal - different

b_{Eng} - p_{Fr}

- phonology - different
- gestural plan - different
- acoustic signal - identical

Planning, and Supervision of Execution

Planning in speech production is about specifying the DYNAMIC SPEECH SCENARIO. In ARTICULATORY PHONOLOGY gestures are marshalled with the timing relationships to ensure that subsequent task dynamic processing results in an acoustic signal enabling perceptual recovery of the gestures *and* the gestural plan.[10] The REVISED MOTOR THEORY[11] is the corresponding theory of speech perception.

The data presented in this paper is not, however, consistent with the notion that the gestural plan can be carried through from its abstract level to the physical articulatory level, allowing simple *non-cognitively based* co articulatory effects to explain why unexpected acoustic signals arise. Although ARTICULATORY PHONOLOGY implies a carry though is possible, it does not provide the basis for explaining the acoustic facts. But because task dynamics cannot modify its procedures, the burden of explanation must rest with ARTICULATORY PHONOLOGY or with an additional external component.

Furthermore, in the TASK DYNAMIC MODEL co articulation is modelled as *gestural layering.*[12] This implies a source external to the model to guide* some coarticulatory processes. Gestural layering (visible in the gestural score - Fig. 3) cannot however exist at the deepest level of planning: placing it there results in a loss of generalisation. The example from our data would be that to represent /p/$_{Eng}$ with gestural layering to explain the subsequent long VOT, and not to represent /p/$_{Fr}$ similarly, would destroy the transparency of the functional similarity of the segments in their respective languages. It also moves a process belonging to the physical stages of speech production into the planning stages - unless it is held that *all* coarticulation is planned. The other approach is to model coarticulation as internal to the task dynamic process, but which *may* be externally supervised.

* The full detail of the action is not specified internally to the model, but actions do have internal, private, characteristics. These characteristics are not externally sourced every time the action is performed.

An early work on coarticulation,[13] based in a less dynamic, segmentally-oriented theory, modelled co articulation as a two-layer process. Using /k/ in English as an example, I claimed that it seemed ‘appropriate to talk of the gesture /k/ on which are superimposed certain co articulation effects due to a predictable physiological constraint’. This type of coarticulation which had been identified earlier[14] was not of linguistic origin and *not planned.* But I go on to identify circumstances in which this language universal coarticulation *can be constrained for use in* a *language-specific way* – and that *is* planning. There is thus a need to model cognitively-sourced supervision of a wholly physical phenomenon. Phonology should not model the detail of this supervision – it is not part of *what* a speaker wants the physical system to do, it is part of *how* the speaker wants the physical system to carry out planned ‘requirements’.

Translating this into contemporary terminology: the TASK DYNAMIC MODEL performs better if in addition to an underlying gestural plan it receives an input from an external component with a supervisory role. The supervisory component is responsible for overseeing the DYNAMIC SPEECH SCENARIO which will unfold under the control of the TASK DYNAMIC MODEL.

This point has been argued very strongly[15] in connection with modelling the causes of observed variations in articulatory precision. We argued that it is not possible to explain *why* precision of articulation varies during the course of utterances simply from the underlying

phonology (in current terms, the gestural score) and a-linguistic coarticulatory phenomena. The coarticulation supervisor was introduced to take account of predictions based on a running model of perception to determine areas of an utterance which required increased articulatory precision. The model of speech production current at the time was the predecessor to the ARTICULATORY PHONOLOGY and TASK DYNAMIC MODEL. The model involved very high level computation of many aspects of motor control which are now (thanks to Fowler[7]) more correctly modelled as properties of low-level coordinative structures. But the proposal that there is more to speech production than the unsupervised execution of a plan holds true.

It would destroy the simplicity of ARTICULATORY PHONOLOGY and the elegance of the TASK DYNAMIC MODEL if we tried to extend in an unprincipled way either of these parts of the overall speech production model to take care of the apparent anomalies of the data cited here. Having different gestural scores for $/p/_{Eng}$ and $/p/_{Fr}$ would obscure the insight that the two objects function similarly in the two languages, and attempting to assign different behaviours to the coordinative structures involved in vocal cord setting for different languages would destroy the important organism-oriented a-linguistic stance of the TASK DYNAMIC MODEL. A supervisory component is needed, separately identified, but incorporated into ARTICULATORY PHONOLOGY.

Gestural Scores

As an illustration of some of the ideas presented here, let us look at the gestural scores which ARTICULATORY PHONOLOGY might propose for the data described earlier. Fig. 3a is how Browman and Goldstein might present the gestural scores for the French *une panne* and *une banne,* and Fig. 3b is what the gestural scores for English *a pan* and *a ban* might look like.

If we compare Figs. 3a and 3b we find that in all four scores by using gestural layering provision is made for velar co articulation resulting in nasality toward the end of the vowels preceding [n] - N. The score representation here is by analogy with scores already given by Browman and Goldstein[5] for English *pan* and *ban,* and is a clear indication that they regard this type of coarticulation as planned into the articulation. On this point I disagree; this is precisely the kind of unsupervised coarticulation which does *not* require information from outside the TASK DYNAMIC MODEL. It is also precisely the type of coarticulation which, because it has no linguistic significance, is *not* recovered later during perceptual processing.

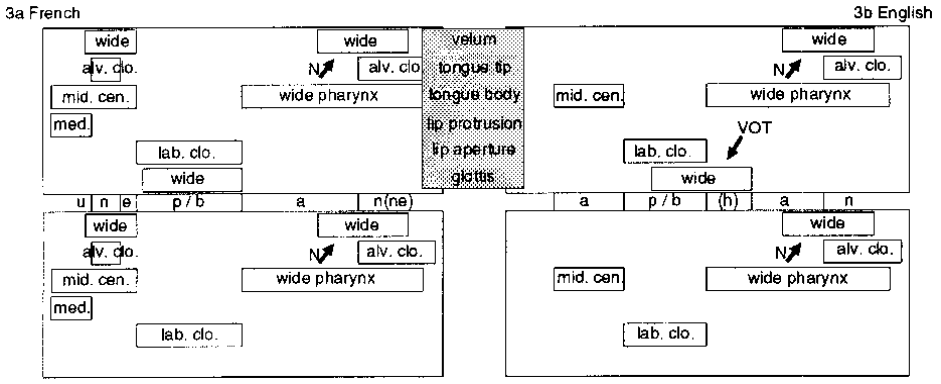

Fig. 3 Gestural scores without supervision.

Note that the scores for *panne* and *pan* differ in that *pan* has glottal opening extending significantly beyond the end of bilabial closure - this is to account for the VOT following the release. Since there is no VOT in French the glottis becomes closed for vocal cord vibration simultaneously with the end of bilabial closure. Thus functional similarity between $/p/_{Eng}$ and $/p/_{Fr}$ becomes opaque and the score cannot capture an important fact of the languages. The score ends up prompting the TASK DYNAMIC MODEL where it is not necessary.

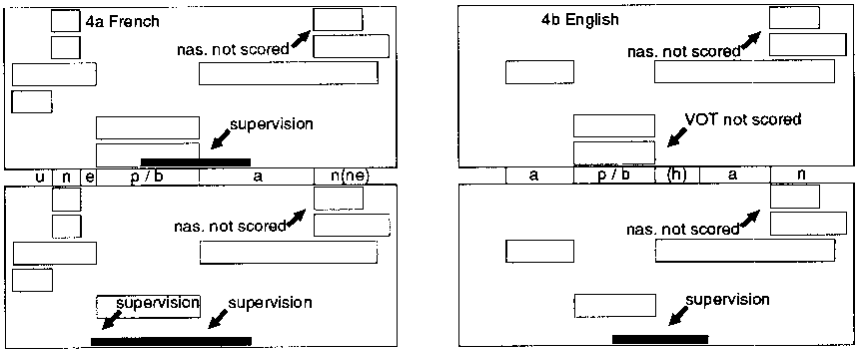

Fig. 4 *une panne, une banne,* a *pan,* a *ban* - supervised gestural scores.

Figs. 4a and 4b illustrate alternative gestural scores, this time marking (with a black bar) where a cognitively driven external supervisor would step in to constrain subsequent errors which would be generated by the TASK DYNAMIC MODEL, and which might potentially result in perceptual failure. Note, though, that the velar co articulation is not included – task dynamics will take care of this automatically: planning is superfluous. In addition the gestural scores for /p/$_{Eng}$ and /p/$_{Fr}$ are shown as identical, as are those for /b/$_{Eng}$ and /b/$_{Fr}$ – thus capturing functional similarity. The solid black bars added to three of the gestural scores indicate where and when the supervisor must step in to exercise local control over the DYNAMIC SPEECH SCENARIO to prevent serious output errors, and to mark constrained co articulation which *is* linguistically significant. This *is* information which will be recovered during perception, so should figure in the gestural score.

No supervision is required (for the glottal opening parameter on which we are focusing) for p_{Eng} – VOT is allowed. But supervision is required for p_{Fr} – VOT must be prevented. Similarly for both b_{Fr} and b_{Eng} VOT must be prevented, so the plan instructs the supervisor to step in. Notice, though, that b_{Fr} is different from b_{Eng} in the supervision of glottal closure during [b]. It is important for French that not only should there be no VOT following the release of the [b], but there must be an attempt to produce vocal cord vibration throughout the stop. In English, however, vocal cord vibration during the stop phase is not of linguistic significance and will not be recovered during perceptual processing – so it can be allowed to fail.

In all cases where supervision is indicated on the score there is a plan to *employ* the supervisor – but the score does not say *how* the supervisor is to operate. The supervisor has cognitive knowledge of how to manipulate the parameters of the TASK DYNAMIC MODEL to produce the necessary effects to maintain phonological integrity and thus optimise the chances of perceptual success. But this knowledge is itself irrelevant to the plan and to the perceptual recovery of significant gestural information.

Conclusion

ARTICULATORY PHONOLOGY and the TASK DYNAMIC MODEL of speech production have significantly advanced our understanding of the relationship between the cognitive and physical aspects of speech production. But they do not handle well the subtle manipulation of constraints within the physical processing, when this manipulation is for clear linguistic purposes. This is not a matter of planning tasks to produce acoustic goals, but of manipulating inbuilt constraints to achieve linguistically significant *task variants.* I have argued in the past that it is important to separate this type of process from the usual phonological processes. And I have argued in this paper that there is something to be gained by adding a cognitive supervisory component to the cognitive planning and physical execution components of the model. This suggestion does not shift the focus of the TASK DYNAMIC MODEL away from the task, but enhances the definition of the task itself.

References

1. J.J. Ohala and J.J. Jaeger (eds.) (1986) *Experimental Phonology.* Academic Press, London

2. C.P. Browman and L. Goldstein (1986) Towards an articulatory phonology. In C. Ewan and J. Anderson (eds.) *Phonology Yearbook* 3. Cambridge University Press, Cambridge, 219-252

3. M.A.A Tatham (1986) Towards a cognitive phonetics. *Journal of Phonetics* 12, 37-47

4. C.P. Browman and L. Goldstein (1992) Articulatory phonology: an overview. *Phonetica* 49, 155-180

5. C.P. Browman and L. Goldstein (1993) Dynamics and articulatory phonology. Haskins Laboratories *Status Reports on Speech Research,* SR-113, 51-62

6. E. Saltzman (1986) Task dynamic coordination of the speech articulators: a preliminary model. In H. Heuer and C. Fromm (eds.) *Generation and modulation of action patterns.* Springer-Verlag, Berlin and Heidelberg, 129-144

7. C.A Fowler, P. Rubin, R.E. Remez and M.T. Turvey (1980) Implications for speech production of a general theory of action. In B. Butterworth (ed.) *Language Production.* Academic Press, New York NY, 373-420

8. M.A.A. Tatham and K. Morton (1973) Electromyographic and intraoral air pressure studies of bilabial stops. *Language and Speech* 16, 335-350

9. K. Harris, G.F. Lysaught and M.M. Schvey (1965) Some aspects of the production of oral and labial stops. *Language and Speech* 8, 135-147

10. C.A Fowler and L.D. Rosenblum (1991) The perception of phonetic gestures. In LG.Mattingly and M. Studdert-Kennedy (eds.) *Modularity and the Motor Theory of Speech Perception.* Lawrence Erlbaum Associates, Hillsdale NJ, 33-59

11. A.M. Liberman and I.G. Mattingly (1985) The motor theory of speech perception revised. *Cognition* 21,1-36

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

13. M.A.A. Tatham (1971) Classifying allophones. *Language and Speech* 14, 140-145

14. S. Ohman (1966) Co articulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America* 39,151-168

15. M.A.A. Tatham and K. Morton (1980) Precision. *Occasional Papers* 23, Linguistics Dept., Essex University, 107-116