Timing in Speech Production with Special Reference to Word Form Encoding

WILLEM J. M. LEVELT
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

INTRODUCTION

The ability to speak is probably our most complex cognitive–motor skill. It is, moreover, a uniquely human and a universal skill. In speaking, myriad processes involving a wide range of cerebral structures cooperate in the generation of a temporally organized structure, an articulatory pattern that has overt speech as its physical–acoustic effect.

The temporal organization of speech is multileveled. There are, on the one hand, the relatively slow strategic processes involved in planning the speech act. When we speak, our attention is almost fully dedicated to *what* we say. *How* we say it largely takes care of itself. Words, for instance, are produced at a speed of about 2 per second, but so-called anacruses are possible of up to 7 words per second. At this rate we retrieve lexical items from a mental lexicon that contains thousands, and probably tens of thousands of items. In fluent speech our average syllabic rate is about 3 per second, whereas individual speech sounds come as fast as 10 to 15 phonemes per second. And normally, all this happens without any attentional control.

These high-speech automatic processes are, moreover, surprisingly error proof in normals. Estimates of the rate of lexical selection errors range around one per thousand, whereas phonemic errors are even rarer. What are the mechanisms that subserve this perfect, multilevel timing in speech production?

In the following I will discuss some recent research in our laboratory that is concerned with the time course of spoken word production at three levels of processing, as depicted in FIGURE 1. The first one concerns lexical selection, the second one phonological encoding and syllabification, and the third one phonetic encoding, in particular the retrieval of syllabic gestural scores.

LEXICAL SELECTION

How do we select the appropriate words for the concepts that we want to express? Ardie Roelofs[1] proposed an activation spreading model for this process. FIGURE 2 presents a fragment of the lexical network.

Lexical items are represented at three levels. An item's meaning is specified at the conceptual level by way of a network of labeled relations. The concept of sheep, for instance, is represented by a conceptual node SHEEP, which entertains an *isa* relation to ANIMAL, etc. The next level is a syntactic stratum. Each lexical concept (such as SHEEP) connects to a so-called *lemma* node at this stratum. Its network

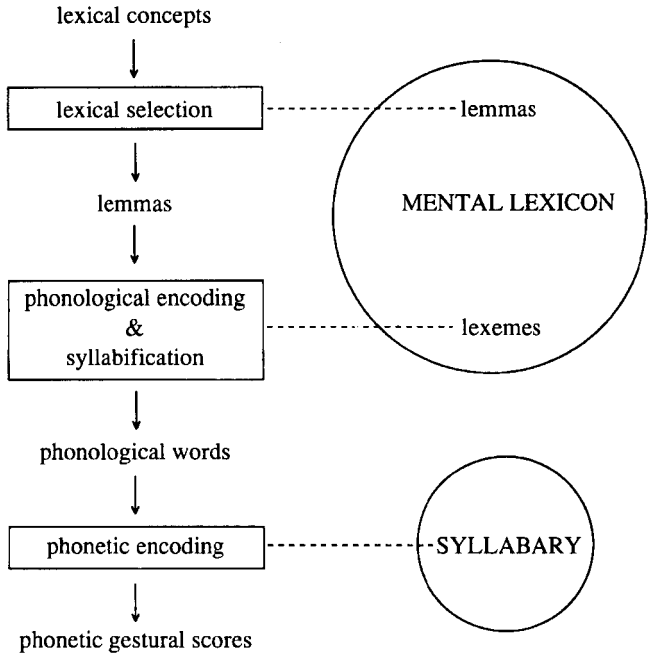


FIGURE 1. Producing words in speech production. Three levels of processing.

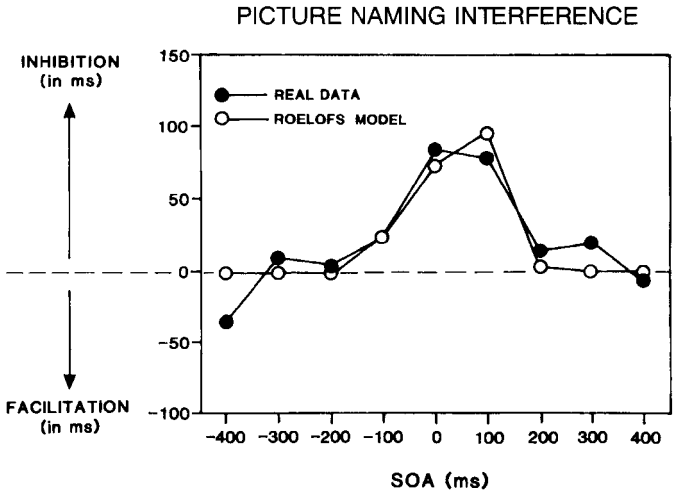


FIGURE 2. Fragment of lexical network. *Arrows* represent types of connections, not the flow of information. (From Bock and Levelt.[2] Reproduced by permission.)

connections at this level represent the item's syntactic properties (for instance that *sheep* is a noun or that French *mouton* has male gender). Finally, there is the lexeme or sound form stratum. Here the item is represented by a *lexeme* node, which in turn connects to segmental and other sound form nodes that specify the item's phonological properties (see below). Each lemma node connects to one lexeme node. But in case of homonyms two different lemma nodes project onto the same lexeme node (see below).

The network has a simple activation spreading regime, which runs in discrete time steps (see Roelofs' original publication for details).

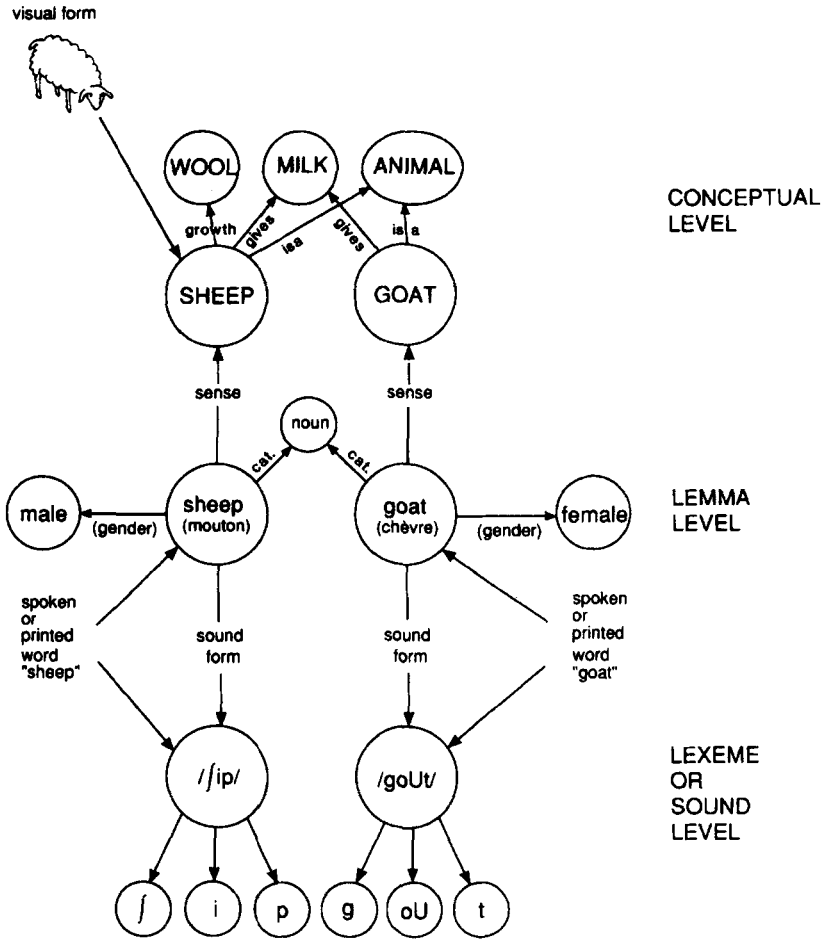


FIGURE 3. Picture naming latency differences for semantically related vs. unrelated primes at nine different SOAs and model simulations. (After Roelofs.[1] Reproduced by permission.)

Roelofs defined lexical selection as the selection of an appropriate lemma node. The time course of lexical selection (for instance, in picture naming) is predicted from a simple probabilistic rule. The probability that a particular lemma is selected during time interval *i* is the ratio of its activation to the sum activation of all active lemmas (the Luce ratio).

The probabilistic character of the rule creates the possibility of explaining errors of lexical selection, such as *nephew* for *uncle.* Given the rule, there is always a small probability that a nontarget item will be selected, such as *goat* instead of *sheep.* When the concept node SHEEP is active, some of its activation will spread to the semantically related concept GOAT, and down to its lemma node *goat.* Hence, errors of selection will often be semantic in character.

But Roelofs tested his (computer-implemented) model by way of reaction time experiments. The basic procedure was to do a picture naming experiment, and to measure the subjects' naming latencies. This process was interfered with by presenting visual prime words that the subject had to ignore. The visual prime could be semantically related to the target word (for instance, "goat" when the picture was one of a sheep), or it could be unrelated. The prime word could be presented at various moments, either before, simultaneous with, or after picture onset (i.e., at different stimulus onset asynchronies or SOAs). The model gave precise predictions for the effect of different types of prime word at different SOAs, and they were surprisingly well confirmed by the experimental data. In addition, the model could account for the major data sets in the literature. FIGURE 3 presents the classic data obtained by Glaser and Düngelhoff[2] and the model's excellent fit.

As soon as a lemma has been selected, it sends its activation down to its lexeme node.

PHONOLOGICAL ENCODING

In phonological encoding we generate the phonological form of an utterance, in particular its segmental and prosodic structure. Central to phonological encoding is the construction of successive syllables, the basic units of articulation. In connected speech syllabification often straddles word boundaries. When we say *Peter gave it,* we contract *gave* and *it* to form a single so-called "phonological word" /geI-vIt/. Here, the syllable boundary ignores the word boundary.

FIGURE 4 diagrams some of the main processes involved in phonological encoding. After a lemma (such as *gave* or *it*) is selected, its lexeme is activated (here FIG. 4 connects to FIG. 3), and two kinds of phonological information become available. The first one is the word's segmental composition, roughly the string of phonemes it consists of. The second one is the word's metrical or foot structure; this is the word's syllabicity (the number of syllables the word contains), and the word's stress pattern over these syllables.

The metrical patterns of successive words will be grouped into (larger) phonological words. And, finally, the "spelled out" string of segments will be associated to a phonological word's metrical frame. This process of association provides, one by one, the successive syllables of which the phonological word is composed.

I now discuss some aspects of the five processes depicted in FIGURE 4, beginning with lexeme activation, and ending with syllabification.

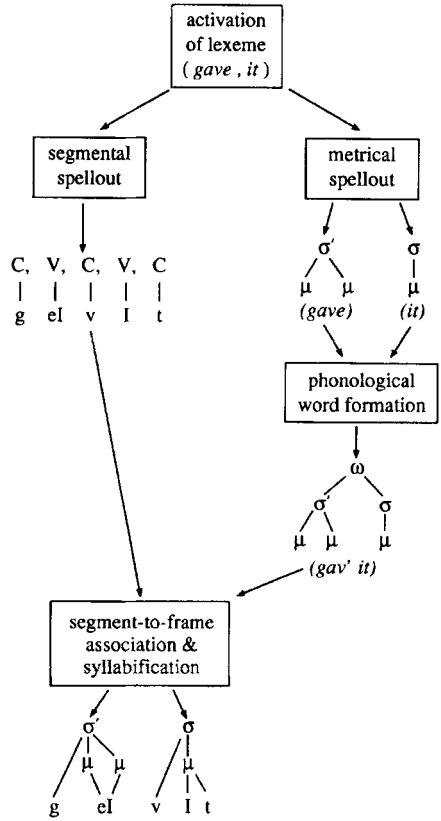


FIGURE 4. Processes involved in the phonological encoding of words.

Lexeme Activation

The proximal cause for a lexeme's activation is the *selection* of its lemma (see FIG. 2). Levelt *et al.*[4] showed experimentally that a merely *activated* lemma does not send its activation to its lexeme node. This state of affairs is different from what one would expect on the basis of existing connectionist and cascading accounts of lexical access, where activation spreads uninterruptedly throughout the lexical network. The mechanism of phonological encoding is apparently carefully sealed from the competitive storms in lexical selection; it has to deal only with the eventual winner, the selected target word.

Another important aspect of lexeme activation is that it is the seat of the word frequency effect in production. Since Oldfield and Wingfield's seminal paper,[5] it is known that in picture naming it takes longer to access low-frequency names than high-frequency names. Wingfield[6] showed that this is not due to recognizing the

depicted objects; the effect is of lexical origin. Jörg Jescheniak and I (in preparation) replicated both findings and then asked ourselves whether the locus of the effect is at the lemma level or at the lexeme level.

In order to test whether it is at the lemma level, we gave our (Dutch) subjects a gender decision task. They were presented with pictures, and on the appearance of each picture they had to push one of two push buttons, corresponding to the gender of the picture's name (is it a *het* or a *de* word?). Gender is a property of lemmas (see FIG. 2). If lemma thresholds are frequency sensitive, then the same frequency effect as in naming should appear in this task. But it did not. After acquainting themselves with stimuli and task, gender decision showed no effect of word frequency, whereas the frequency effect in naming appears fully fledged for the same pictures, and cannot even be eradicated by repeated presentations of the pictures.

That the effect is indeed due to accessing the word's sound form could be shown in an experiment where subjects were asked to name the low-frequent item of a homonym pair. For instance, they would name the animal *bee,* where there is a high-frequent homophone *be.* Our network theory has different lemma nodes for these two items (one is a noun, the other a verb). But they project onto one and the same lexeme node, /bi/. Therefore *bee* should behave like a high-frequent item rather than like a low-frequent one, and that is what we found.

The conclusion here is that a lexeme's threshold activation, that is, the activation needed for releasing its phonological information, is frequency dependent. In normal speech, the release of this information is occasionally (but rarely) blocked, leading to the much studied "tip-of-the-tongue" phenomenon (see Levelt[7] and Brown[8] for reviews). The same mechanism is involved in the pathological case of anomia.

Segmental Spellout

Since the beginning of speech error research it has been known that a word's segments can be independently affected or displaced in spontaneous speech. A spoonerism such as *With this wing I thee red,* shows that a word's phonological form does not become available as an indivisible template. Rather, phonemes are independently released and positioned into some independently generated metrical word or syllable frame (Shattuck-Hufnagel[9]). But speech errors do not tell us whether a word's segments are simultaneously, or rather successively released. What is the timing of segmental spellout?

Following up on initial findings by Antje Meyer[10, 11] that suggested successive spellout of segments, Meyer and Schriefers[12] used the priming technique to measure the time course of phonological spellout. In a picture naming task they presented their subjects with prime words that were phonologically related to the picture's name or with unrelated control words. For instance, the picture could be one of a cigar (Dutch name: *sigaar,* pronounced [si-xa:r]). The prime word could be *citroen* ([si-tru:n], which shares the first syllable of the target (begin-related prime). Or it could be *bulgaar* ([bül-xa:r]), which shares the second syllable of the target (end-related prime). As a control prime, a phonologically unrelated word was used, such as *boutique.*

The prime could be presented such that its related syllables ([si] and [xa:r], respectively, for begin- and end-related primes) began at either 300 or 150 ms before the picture, simultaneously with the picture, or 150 ms after picture onset. The

subjects were instructed to ignore the prime and to name the picture as soon as it appeared. Naming latencies were measured.

The central finding was this: At SOA = −300 ms neither of the two primes had any significant effect on the naming latencies (as compared to the controls). The begin-related primes, however, began facilitating the response at SOA = −150 ms, where end-related primes were still without effect. The facilitatory effect of end-related primes began 150 ms later, at SOA = 0. This shows that a word is not phonologically encoded as a whole, but incrementally from beginning to end. Antje Meyer and Herbert Schriefers[12] could show that the same holds for monosyllabic words.

Linda Wheeldon and I (in preparation) obtained rather precise data on the time course of phonological spellout by way of a quite different technique. We replaced the usual picture naming task by a *translation* task. Here (Dutch) subjects were given a list of English–Dutch translation equivalents, for instance *hitch-hiker–lifter.* Since all subjects knew (some) English, each word's translation was easily memorized. As soon as this was the case, we introduced the experimental task. The subject was given a phoneme target, for instance /f/, and instructed to push a yes button every time the Dutch translation of a new English word on the screen contained that target. Hence, the subject pushed the yes button shortly after the word *hitch-hiker* was presented on the screen; this is because *lifter* contains an /f/. And, of course, we measured the response latencies. Notice that subjects did not utter the Dutch translation words; they only performed their phoneme monitoring.

It is likely that this task directly measures the timing of segmental spellout. If a word like *lifter* is spelled out "from left to right," then its consonantal phonemes /l/, /f/, /t/, and /r/ will become available one after another, and this should affect the monitoring latencies. When /l/ is the target phoneme, monitoring should be relatively fast, and it should be increasingly slower when /f/, /t/, and /r/ are the target phonemes.

The experiment was run over 20 items, all with CVCCVC structure like *lifter,* where each of the consonants involved could be a target. The results confirmed the expectations. The monitoring latency was 1178 ms on average for the first consonant (i.e., for /l/ in *lifter*), 1233 ms for the second consonant (i.e., for /f/), 1289 ms for the third consonant (/t/ in the example), and 1302 ms for the final consonant (/r/ in *lifter*). We are confident that the subjects are not monitoring their internal speech in this experiment. The results are essentially the same when subjects are given a concurrent counting aloud task during the experiment.

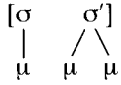
It is interesting to consider the size of these significant increases. The first and third consonant are exactly one syllable apart, their latency difference is 111 ms. The second and fourth consonant are also one syllable apart, their latency difference is 69 ms. The average duration of a spoken syllable is about 250 to 350 ms. Apparently, the speed of spelling out is two or three times as fast as the speed of articulation.

It is an important question what it is that is spelled out. The segments are probably not fully specified. Stemberger[13] argued this point on the basis of speech errors such as *in your really gruffy–scruffy clothes.* In *scruffy* the second segment (/k/) is probably unspecified for voicing. It will acquire the correct feature (−voiced) at a later stage in the process (see below); in English +voiced is impossible in the context s−r. But in the error, where /s/ is lost, the context (−r) is insufficient to provide the feature specification, and it may then happen that the underspecified segment surfaces as /g/ instead of /k/. It is fully in line with modern phonology (cf. Archangeli[14]) to suppose that a word's segments are stored and spelled out in underspecified form.

It is, in fact, better to reverse terminology here. It is not so much segments that are spelled out, but small sets of feature specifications. The second segment of *scruffy,* for instance, is probably only specified as +velar, no more. Segmental spellout, then, is the retrieval of these minimal feature specifications for successive "timing slots."[a]

Metrical Spellout

When speakers are in a tip-of-the-tongue state, they can often report on the number of syllables and the stress pattern of the trouble word. This suggests that metrical spellout can proceed independently of segmental spellout. In Levelt,[15] I proposed that (for English) the metrical spellout of a word consists of the number of syllables, their weights, and stress pattern. For the word *neglect* it would be



where the first syllable is light (one μ) and the second one is heavy (two μ) and accented.

If metrical structure is independently represented, it should be possible to prime its spellout, independent of the word's segmental composition. Paul Meyer and I (in preparation) could show that this is indeed the case. In one experiment we used the priming procedure that Antje Meyer and Herbert Schriefers[12] had used (see above). The subjects had to name pictures that all had two-syllable names. For half of the pictures the name had iambic meter (such as [si-xa:r], similar to *cigar* in English); for the other half, the meter was trochaic (such as [moU:tər], like *motor* in English). For each picture subjects heard a disyllabic prime word that they had to ignore. The prime word could be presented at different SOAs, but here I will ignore that variable. The experimental variable was whether the prime word corresponded in meter to the target word, and we measured subjects' naming latencies.

The results of this experiment were clear. We obtained a highly significant 58-ms facilitation effect when the prime had the same meter as the target, but this occurred under one condition only: the first segment of prime and target had to be identical. For instance, *saloon* is a better prime for *cigar* than is *salmon,* but *balloon* and *ballot* are equally ineffective. This effect had been predicted by Paul Meyer. If segmental and metrical spellout run in parallel, as is suggested in FIGURE 4, priming metrical spellout will only be effective if it is the slowest of the two processes. In the effective condition segmental spellout is given a head start; segmental spellout is facilitated by the word-initial identity of prime and target.

These findings could be replicated by using the translation task as experimental procedure (see above). Here the subjects produced the Dutch translation of an English word on the screen, while they heard an acoustic metrical prime that they had to ignore.

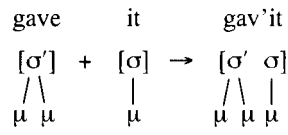
Together, these results form the first reported experimental evidence for the independent generation of a word's metrical form.

[a]The timing slots are probably also specified as C (consonantal) or V (vocalic or sonorant), as indicated in FIGURE 4. A word can namely be primed by another word of the same CV-composition, as Paul Meyer and I (in preparation) could recently show.

Phonological Word Formation

Any utterance has a multilevel prosodic structure. At the top level there are intonational phrases, defined by a characteristic pitch contour. Intonational phrases consist of phonological phrases. These are metrical phrases that have lexical heads-of-phrase as their final elements (as in *The committee / had considered / that the students / might have needed / personal computers /*). In their turn, phonological phrases consist of phonological words. In the example, the phonological phrase *might have needed* consists of two phonological words, *might've* and *needed.* At all three levels metrical planning is sensitive to syntax (see Levelt[7] for a review). Here I will only consider the formation of phonological words.

A major process in phonological word formation is *encliticization.* Here a light lexical element is attached to a preceding head word, like in *might've,* or *gav'it.* This process is sensitive to syntax. Encliticization is blocked when there is a major syntactic boundary between the two elements (one cannot cliticize *it* to *gave* in *What Peter gave, it should be stressed, is irrelevant*). But though phonological word formation depends on syntax and on the metrical composition of the lexical elements involved, it is independent of the segmental composition of these elements. Hence, one can characterize phonological word formation as a purely metrical process. The formation of *gav'it,* for instance, can be formally represented as



(1)

Although the rules of cliticization and phonological word formation are rather well-understood (see, for instance, Nespor and Vogel[16]), literally nothing is known about the implementation of these rules in the *process* of phonological word formation as it develops over time.

Segment-to-Frame Association and Syllabification

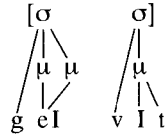
The final stage of phonological encoding consists of associating the string of spelled-out segments to the phonological word's metrical frame. Above I mentioned Paul Meyer's finding that for metrical priming to appear, segmental spellout should be given a head start. This indicates that metrical spellout is relatively fast. Normally, the metrical frame is already there to absorb successive segments as they are spelled out. Levelt[15] proposed that segments are, one by one, attached to the metrical frame, going "from left to right," so to say. The following rules of attachment (for English) were proposed in that paper (still excluding the diphthong rule):

(i) A vowel only associates to μ; a diphthong to μμ.
(ii) The default association of a consonant is to σ. A consonant associates to μ if and only if any of the following conditions hold:

—the next element is lower in sonority;
—there is no σ to associate to;
—associating to σ would leave a μ without an associated element.

(In addition there is the general convention that attachment to σ can only occur to the left of a syllable's morae).

For the rationale of these rules I refer the reader to the original paper. Here I will, by way of example, apply the rules to the generation of the phonological word *gav'it.* The spelled-out segments /g/, /eI/, /v/, /I/, and /t/ are successively attached to the right-hand structure in (1). The first segment, /g/, is a consonant and has to attach to σ, according to rule (ii). The second segment is the diphthong /eI/, which attaches to μμ, according to rule (i). The third segment is /v/. According to rule (ii) it must associate to σ, but that can only be done to the left of a syllable's morae. Hence, the association has to go to the next σ, inducing a syllable break. The fourth segment /I/ attaches to μ according to rule (i). And the fifth segment /t/ will attach to the same μ because the second condition under rule (ii) holds. The final result is



It is important to notice that syllabification takes place "on the fly" as successive segments are attached to the metrical frame. Different from what standard terminology in phonology suggest, there is no *resyllabification.* It is not the case that a word's segmental syllable composition is stored, retrieved, and subsequently changed (resyllabified) as phonological words are formed. That would be a wasteful process. Rather, the independent spellout of segmental and metrical information makes it possible that phonological word formation runs on metrical information only. Syllabification then comes "for free" at the later stage of segment-to-frame association.

The eventual output of phonological encoding is a metrically structured string of phonological syllables. If, as I suggested earlier, phonological segments are underspecified, then these phonological syllables are underspecified as well. How, then, does the speaker compute the full phonetic form of each syllable? This brings us, finally, to phonetic encoding (see FIG. 1).

PHONETIC ENCODING

Phonetic encoding involves the production of what Browman and Goldstein[17] have called a *gestural score.* A gestural score is a specification of the gestural "tasks" that have to be performed over time by the various articulatory subsystems in order to produce the target utterance. According to Browman and Goldstein there are five subsystems whose gestures can be independently controlled. Hence there are five "tiers" in a gestural score. They are the glottal and the velar system, and three tiers in the oral system.

TABLE 1 represents the tasks that can be specified for each of these subsystems. At the lips tier, for instance, the task is two-dimensional. Lips can be instructed to protrude, and they can be instructed to open or close.

Normally, a word's articulatory gesture results from performing tasks at different tiers simultaneously. But a task underspecifies the gesture. Take, for instance, the task of closing the lips. It can be realized in infinitely many ways. One can move the

TABLE 1. Gestural Tasks on Articulatory Tiers

Tier	Task Variables
Glottal	Aperture
Velar	Amount of closure
Oral—Tongue body	Place and amount of constriction
—Tongue tip	Place and amount of constriction
—Lips	Protrusion and aperture

upper lip, the lower lip, the jaw, or all three of them to different degrees. Which combination will be used by the speaker depends on myriad circumstances, such as the starting position of the articulators or arbitrary physical contingencies (e.g., having a pipe in the mouth).

How the articulatory system factually executes a particular gestural task is a fascinating problem in coordinative structures theory (Saltzman and Kelso[18]), but it is not the topic of phonetic encoding. Our problem is "merely": Where do gestural scores come from? There are two approaches here, which are not mutually exclusive, I believe, but rather complementary.

The first one is the direct route. The idea is that a word's phonological specification is already an abstract rendering of its gestural score. The features in successive timing slots are essentially specifications of phonological tasks; for instance, that there should be velar closure at some early moment in the word *scruffy*. A sophisticated rendering of this direct route can be found in the work by Browman and Goldstein.[17]

Here I would like to argue for a more indirect route. I suppose that speakers have access to a *mental syllabary*. This is a repository of phonetic programs or gestural scores for the syllables in the speaker's language. As phonological syllables are generated one after another (see above), they will function as access codes to the syllabary. Each of them will trigger the retrieval of the corresponding gestural score, which in turn will be executed by the articulatory system.

One argument for the existence of a syllabary is that syllables are real units of articulation; within-syllable phonetic coherence is much larger than between-syllable coherence. Moreover, most syllables are highly overused units of articulation. It would be wasteful to fully program them time and again.

The syllabary theory is, of course, more attractive for languages such as Chinese and Japanese, where the number of syllables is no more than a few hundred, than for English, which has some 6,000–7,000 different syllable patterns. But even for English the amount is not excessive; the number of words the speaker has in store is very much larger.

An obvious advantage of the syllabary theory is that phonological underspecification becomes an almost trivial problem. There is no need to "complete" the specifications of successive segments in a word. The only condition that has to be satisfied in the syllabary is that each phonological syllable (consisting of underspecified segments) corresponds to one and only one gestural score. That score, then, is fully specified in the syllabary.

One nontrivial prediction from the syllabary theory is that there should be a frequency effect. Just as low-frequent words are harder to access than high-frequent ones (see above), low-frequent syllables should be harder to access than high-

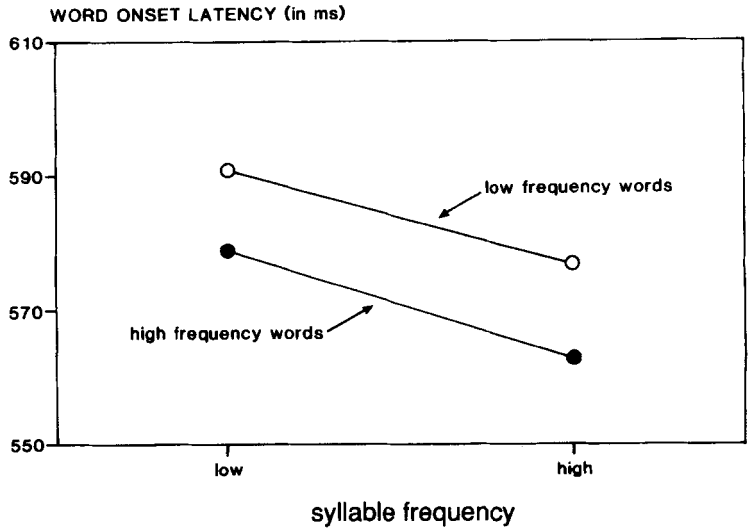


FIGURE 5. Naming latencies for high- and low-frequency words consisting of high- or low-frequency syllables.

frequent ones. Moreover, these two effects should be independent, because lexicon and syllabary are independent stores.

In order to test this theory, Linda Wheeldon and I (in preparation) had people produce two-syllable words (in response to abstract visual patterns that they had learned to associate with these words). We used four types of words. There were low-frequent words consisting of low-frequent syllables (such as *lantern*), low-frequent words consisting of high-frequent syllables (such as *litter*), high-frequent words that consisted of low-frequent syllables (such as *language*), and high-frequent words consisting of high-frequent syllables (such as *lady*). The response latencies obtained are presented in FIGURE 5. The results are as predicted; there is both a word and a syllable frequency effect, and the two effects are independent (or additive). Although these findings are open to alternative explanations, they do invite further exploration of the syllabary notion.

CONCLUSION

The present paper outlined some of the major steps involved in spoken word production. I reviewed a research program that analyzes each step by experimental procedures specifically affecting or tapping into its time course. As the partitioning of this complex system becomes more transparent, further questions can be profitably raised. Among them are issues in language and speech pathology, such as the origins of disturbances of lexical selection, anomias, and disorders of timing in word formation. Also one can, with some confidence, begin to relate various component processes in the model to specialized cerebral structures by making use of

brain scanning imagery in combination with experimental procedures of the sort described in this paper.

But it will still be a major step to the analysis of larger stretches of connected speech. Response latencies in normal picture naming are around 600 ms. But we speak at a rate of two to three words per second. Clearly, there is substantial overlap in accessing successive words. It is still largely an enigma how this parallel lexical processing is organized in the brain.

REFERENCES

1. ROELOFS, A. 1992. Cognition **42:** 107–142.
2. GLASER, W. R. & F.-J. DÜNGELHOFF. 1984. J. Exp. Psychol. **10:** 640–654.
3. BOCK, J. K. & W. J. M. LEVELT. 1993. Language production: Grammatical encoding. *In* Handbook of Psycholinguistics, M. Gernsbacher, Ed. Academic Press. New York.
4. LEVELT, W. J. M., H. SCHRIEFERS, D. VORBERG, A. S. MEYER, T. PECHMANN & J. HAVINGA. 1991. Psych. Rev. **98:** 122–142.
5. OLDFIELD, R. C. & A. WINGFIELD, 1965. Q. J. Exp. Psychol. **17:** 273–281.
6. WINGFIELD, A. 1968. Am. J. Psychol. **81:** 226–234.
7. LEVELT, W. J. M. 1989. Speaking: From Intention to Articulation. MIT Press. Cambridge, Mass.
8. BROWN, A. S. 1991. Psychol. Bull. **109:** 204–223.
9. SHATTUCK-HUFNAGEL, S. Speech errors as evidence for a serial order mechanism in sentence production. *In* Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett; W. E. Cooper and E. C. T. Walker, Eds.: 295–346. Erlbaum. Hillsdale, N.J.
10. MEYER, A. S. 1990. J. Mem. Lang. **29:** 524–545.
11. ———. 1991. J. Mem. Lang. **30:** 69–89.
12. MEYER, A. S. & H. SCHRIEFERS. 1991. J. Exp. Psychol.: LMG **17:** 1146–1160.
13. STEMBERGER, J. P. 1982. Lingua **56:** 43–65.
14. ARCHANGELI, D. 1988. Phonology **5:** 183–207.
15. LEVELT, W. J. M. 1992. Cognition **42:** 1–22.
16. NESPOR, M. & I. VOGEL. 1986. Prosodic Phonology. Foris. Dordrecht, the Netherlands.
17. BROWMAN, C. P. & L. GOLDSTEIN. 1991. Gestural structures: Distinctiveness, phonological processes, and historical change. *In* Modularity and the Motor Theory of Speech Perception, I. G. Mattingly and M. Studdert-Kennedy, Eds.: 313–338. Erlbaum. Hillsdale, N.J.
18. SALTZMAN, E. & J. A. S. KELSO. 1987. Psychol. Rev. **94:** 84–106.