intelligibility of nasals and fricatives, at a relatively small cost in naturalness.

A few years after the 1976 transfer of code to Telesensory Systems, Klatt used the Hunnicutt (1980) letter-to-phoneme rules in the design of a complete text-to-speech system, known as Klattalk (Klatt, 1981, 1982a). The system included a 6000-word exceptions dictionary for common words that failed letter-to-sound conversion, and a crude parser. Klattalk software was then licensed to Digital Equipment Corporation in 1982. Digital announced the DECtalk commercial text-to-speech system in 1983 (Bruckert *et al.*, 1983). In designing DECtalk hardware, Digital engineers included sufficient power and flexibility to be able to plug in improved versions of the Klattalk code as they became available in succeeding years (Conroy *et al.*, 1986) (example 33 of the Appendix).

The most recent version of the Klattalk program includes rules to implement such phonetic details as schwa offglides for lax vowels, nasalization of vowels (the splitting of $F1$ into a pole–zero–pole complex) adjacent to nasal consonants, postvocalic allophones for sonorant consonants, variations in voice onset time as a function of syllable structure and stress, target undershoot for short segments (Lindblom, 1963), vowel–vowel coarticulation across an intervening consonant (Öhman, 1966), and breathy offsets to utterances.

Several different voices are provided in Klattalk to approximate the speaking characteristics of men, women, and children. Detailed formant data are stored for only two voices, a man's and a woman's; other male and female voices are created by scaling formant frequencies for different vocal tract sizes and by adjusting an extensive set of synthesis parameters concerned with the voicing source. However, in spite of an ability to modify average f_0, f_0 range, spectral tilt, glottal open quotient, and breathiness, a truly feminine voice quality remains elusive (example 35 of the Appendix). The DECtalk implementation of Klattalk permits the user to modify characteristics of eight preset voices (Conroy *et al.*, 1986).

Apparently oblivious to all of the prior research detailed earlier, a man experimenting in his basement workshop, Richard Gagnon, designed a synthesis-by-rule program that eventually resulted in the Votrax SC-01 chip (Gagnon, 1978; Bassak, 1980). The chip has been interfaced with the Elovitz *et al.* (1976) text-to-phoneme rules (Morris, 1979) and used in several inexpensive text-to-speech products (Sherwood, 1979), the best known of which is the Votrax Type-n-Talk. It is a remarkable device for the price. The chip includes both a cascade formant synthesizer and simple low-pass smoothing circuits for generating continuous time functions to control the synthesizer from a step-function representation derived from target values stored in tables for each phoneme of a somewhat nonstandard phonetic inventory. The latest version of the chip, the SC-1A is used in the Votrax Personal Speech System (example 28 of the Appendix). The new chip is said to have improved intelligibility over the SC-01, but the intelligibility is not nearly as good as obtained in the other systems, and sentence-level rules for prosody and phonetic recoding are not as extensive (see performance evaluation section below).

Another type of chip, the Texas Instrument's TMS-5220 linear prediction synthesizer, forms the basis for a second inexpensive product, the *Echo* text-to-speech system (example 29 of the Appendix). This system appears to use concatenated diphones obtained by excising chunks from natural speech (Peterson *et al.*, 1958; Dixon and Maxey, 1968; Olive, 1977), see below.

A noteworthy early commercial system, the Kurzweil reading machine for the blind, was announced as a product in 1976 (Kurzweil, 1976). It is reputed to have an excellent multifont text reading capability. While admirable in its aspirations, the speech produced by the first versions of this device, which employed phonemic synthesis strategies based on Votrax, was only marginally intelligible (example 27 of the Appendix). Kurzweil currently uses the Prose-2000 as the synthesis hardware in its reading machines.

2. Articulation-based rule programs

A synthesis-by-rule program that manipulates parameters such as formant frequencies according to heuristic rules is not a very close model of the way that people speak. In the hope that a more realistic articulatory model might lead to simpler more elegant rules, several research groups have attempted to devise simplified models of the articulators or models of the observed shape of the vocal tract. The first such model (Kelly and Lockbaum, 1962) used stored tables of area functions (cross-sectional area of the vocal tract from larynx to lips) for each phonetic segment and a linear interpolation scheme. The authors had begun to assemble a list of special case exceptions needed to make this type of strategy work better, such as not constraining the vocal tract except at the lip section when synthesizing a labial stop, and including separate shapes for velars before front and back vowels. Still, the intelligibility of the synthesis was said to be not nearly as good as Kelly and Gerstman had obtained with a formant-based rule program (unfortunately, I have been unable to locate a recording of this system).

Based on the success of Stevens and House (1955) in developing a three-parameter description of vocal tract shapes capable of describing English vowels, the next more ambitious articulatory models abandoned direct specification of an area function in favor of an intermediate model possessing a small set of movable structures corresponding to the tongue, jaw, lips, velum, and larynx. Rules for converting phonetic representations to signals for the control of the position of quasi-independent articulators in an articulatory synthesizer were then developed in several laboratories (Nakata and Mitsuoka, 1965; Henke, 1967; Coker, 1968; Werner and Haggard, 1969; Mermelstein, 1973). The Coker rules were demonstrated at the 1967 M.I.T. Conference on Speech Communication and Processing (example 19 of the Appendix).

Coker found the system to be challenging to work with. For example, in his model shown in Fig. 22, the tongue body position was relative to jaw opening, and the location of the tongue tip was relative to the computed coordinates of the tongue body. If the objective were to make a narrow constriction for, e.g., /s/, several semi-independent articulators

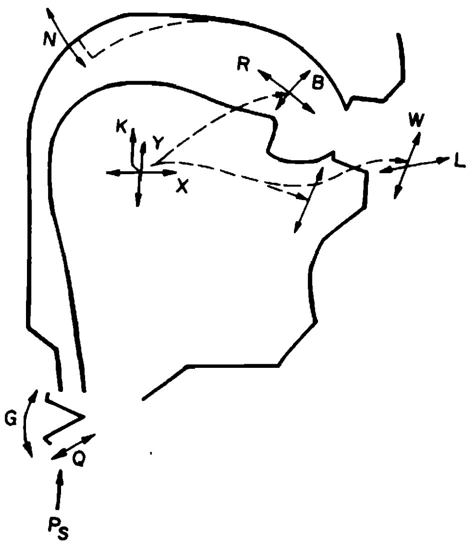

FIG. 22. A simplified low-dimensionality model of the vocal tract (Coker, 1976) permits the lips to open (W) or protrude (L); the tongue body to be raised (Y) or be backed (X) or be bunched for a velar closure (K); the tongue tip to be raised with respect to the body (B) or moved forward (R); the jaw to be raised, and the velum to open a path to the nose (N). The sound generating properties of the larynx are controlled by subglottal pressure (P), static glottal opening (G), and vocal cord stiffness (Q).

(jaw, tongue body relative to jaw, and tongue tip relative to tongue body) all had to be sent to appropriate targets at times that took into account their relative masses and available muscular forces (Coker, 1976). Modern three-dimensional models of the articulators now solve this particular problem of control precision and coordination by grooving the tongue at the midline before forcing it up against the roof of the mouth (Fujimura and Kakita, 1979). However, a general solution to the problem of seeking target articulatory shapes via sets of dependent articulators seems to require control strategies incorporating considerable knowledge of the dynamic constraints on the system and selection of an optimal control strategy from a multiplicity of alternative ways to achieve a desired goal.

Several novel articulation-based synthesis-by-rule programs were developed at this time. Nakata and Mitsuoka (1965) attempted to implement the idea that an intervocalic consonant is a gesture superimposed on an underlying vowel–vowel transition. Henke (1967) proposed an articulatory strategy in which articulators not constrained by the present segmental configurational goals are free to look ahead and begin to seek articulatory goals of upcoming segments. In this way, anticipatory lip rounding and other segmental interactions might be explained on general principles. There is currently considerable disagreement as to the extent to which articulators are free to participate in such lookahead strategies, and as to the number of segments over which lookahead is possible. Finally, Hiki (1970) simulated the muscular control of the articulators in order to be able to specify articulation in terms of motor control signals. This would be a very attractive model if it were the case that the motor commands for a segment were invariant with phonetic context, but unfortunately, electromyographic data indicate that this is far from the case (MacNeilage and DeClerk, 1969).

An entire text-to-speech system for English based on an articulatory model was created in Japan (Teranishi and Umeda, 1968; Matsui *et al.*, 1968) (example 24 of the Appendix). The text analysis and pause assignment rules of this system were based on a sophisticated parser (Umeda and Teranishi, 1975). Using a dictionary of 1500 common words found useful for parsing, the program checked each sentence for length; if it was greater than about ten syllables, it was subdivided into smaller "breath groups" separated by pauses. Some of these rules were later modified slightly and combined with the Coker articulatory rules to produce a text-to-speech system at Bell Laboratories (Coker *et al.*, 1973; Umeda, 1976). The Bell Labs system was notable for its attention to detail in the specification of segmental durations and allophonic variation (example 25 of the Appendix).

While it is possible to generate fairly natural sounding speech using a modern articulatory synthesizer (Flanagan *et al.*, 1975; Flanagan and Ishizaka, 1976, 1978), *rule-based* articulatory synthesis programs have been difficult to optimize. This seems to be due in part to the unavailability of sufficient data on the motions of the articulators during speech production. Even so, the strategies developed to control such a synthesizer may reveal interesting aspects of articulatory control during the production of natural speech (Mermelstein, 1973; Coker, 1976).

3. Rule compilers

Carlson and Granström (1975, 1976) developed a special programming language to permit linguists to formulate synthesis rules in a natural way, similar to the Chomsky and Halle (1968) formalism. An important advantage of the language is an ability to refer to natural sets of phonemes through a distinctive feature notation, making rule statement simple, efficient, and easy to read. These rules are then compiled automatically into a synthesis-by-rule program. A number of languages (Swedish, Norwegian, American English, British English, Spanish, French, German, and Italian) have been synthesized using this system (Carlson and Granström, 1976; Carlson *et al.*, 1982a), and the resulting system has been brought out as a product, the Infovox SA-101 (example 31 of the Appendix). A similar approach has been developed by Hertz (1982), who has used her programming facility to synthesize English and Japanese.

Hertz *et al.* (1985) believe that powerful new rule compilers are needed in text-to-speech systems in order to take advantage of recently proposed linguistic structures such as "three-dimensional" phonology (Halle and Vergnaud, 1980; Clements, 1985). Programmers of synthesis-by-rule systems have always faced the problem that the abstract representation for a sentence is not simply a linear string of symbols. Some rules want to manipulate phonetic segments (while ignoring stress and syntactic symbols), while other rules have a domain that is closer to syllables (or syllable onsets and rhymes), and other rules deal with whole words and phrases. One solution has been to order rules so that it is possible to erase syntactic structure after all syntactic rules have been applied, and erase stress marks after all stress rules have been applied, etc. An alternative, analogous to three-dimensional phonology, is to maintain all forms of representation in parallel (Halle, 1985).

In one sense, rule compilers are an answer to the problem that rule programs written in conventional programming languages nearly always attain a rigidity and opacity that eventually prohibits their developers from making improvements. Rule compilers discourage *ad hoc* fixes and encourage distinctions between levels of description. Indirect support for this view comes from my own work. I have twice found it necessary to re-program the Klattalk text-to-speech system from scratch within a slightly new conceptualization, using a better programming language each time. Nevertheless, I view existing rule compilers as somewhat constraining compared with general programming languages such as "C," and so thus far I have resisted the temptation to make use of them.

A second advantage of rule compilers is the ability to develop a text-to-speech system for a *new* language much more rapidly than when language-specific code and general synthesis strategies are intertwined. This is clearly true when a new team of researchers wishes to build from an existing system (as evidenced by the difficulties that both Speech Plus and Digital Equipment Corporation have had in subcontracting software modification efforts to create systems for other languages), but this need not be the case when the system is well understood (Klatt and Aoki, 1984).

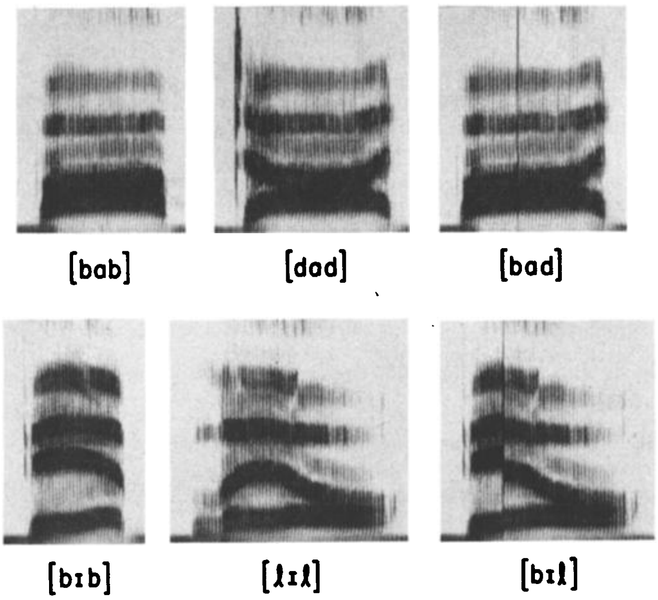

FIG. 23. Broadband spectrograms suggesting that the diphones "ba" and "ad" obtained from the syllables [bɑb] and [dɑd] can be juxtaposed to synthesize a good approximation to [bɑd] (upper right panel). However, synthesizing [bɪl] (lower right panel) from diphones extracted from [bɪb] and [lɪl] requires special care to avoid perceptually disruptive formant discontinuities, see text.

4. Concatenation systems

Other laboratory synthesis-by-rule programs include several that attempt to take pieces of natural speech as building blocks to reconstitute an arbitrary utterance. The recorded chunks cannot be whole words because of the reasons identified earlier. However, smaller units might work.

The syllable is a linguistically appealing unit, but there are over 10 000 different syllables in English. The phoneme is another linguistically well-motivated unit, of which there are about 40 in English. However, all efforts to string together phoneme-sized chunks of speech have failed because of the well-known coarticulatory effects between adjacent phonemes that cause substantial changes to the acoustic manifestations of a phoneme depending on context (Harris, 1953). Coarticulatory influences tend to be minimal at the acoustic center of a phoneme, which prompted Peterson *et al.* (1958) to propose the "*diphone*," i.e., the acoustic chunk from the middle of one phoneme to the middle of the next phoneme, as a more satisfactory unit, Fig. 23.[5] There are thus about 40 times 40, or 1600, different diphone possibilities, although not all occur (Peterson *et al.*, 1958; Sivertsen, 1961). It may be necessary to include several different versions of each diphone to handle distinctions between stressed and unstressed syllables, to include allophones that can occur in different structural environments, and perhaps to include some larger VCV units which Sivertsen (1961) called syllable dyads. In addition, one must be able to change the duration and fundamental frequency contour on a diphone, or perhaps store multiple variants of each diphone with differing prosody. Wang and Peterson (1958) estimated that as many as 8000 diphones may be necessary, but current systems seem able to function with an inventory of about 1000 diphones.

In order to illustrate the advantages of the diphone approach over synthesis-by-rule programs, consider the task of plosive–vowel synthesis. In the rule programs described above, simple theories were used to generate a plosive before different vowels. In the diphone approach, each plosive–vowel transition is a special case, so no general theory or list of exceptions are required.

A potential *disadvantage* of the diphone approach is that discontinuities may appear right in the middle of vowels if the two abutting diphones do not reach the same vowel target, as might be the case for, e.g., the word "bill" in the lower panel of Fig. 23, or for "wet" = [wε + ε^{ə}t] because the [w] lip rounding and velarization effects can extend well into the vowel. Some sort of smoothing at diphone boundaries minimizes the perceptual consequences of actual formant discontinuities, but a mismatch of vowel quality between the two halves is not as easy to compensate for. Nor is it possible to create vowel–vowel coarticulation across an intervening consonant, or adjust vowel targets according to stress or phonetic environment. These may be second-order effects of less importance than a segmental intelligibility gain achieved by diphone concatenation, but we simply do not know.

Efforts to build synthesis-by-rule programs based on the diphone have had considerable success (Dixon and Maxey, 1968; Olive, 1977). The first diphone system, demonstrated at the 1967 M.I.T. Conference on Speech Communication and Processing, was based on a set of stylized stored parameter tracks to control a formant synthesizer (Dixon and Maxey, 1968). The authors spent many years in a trial-and-error effort to optimize a diphone inventory for this purpose (Estes *et al.*, 1964), and eventually produced a system that seemed quite intelligible (example 18 of the Appendix), but the project was terminated for business rather than technical reasons before they were able to add rules for automatically generating segment durations and an f_0 contour from an abstract phonemic representation.

The advent of linear prediction speech analysis/resynthesis techniques opened up the possibility of automated procedures for creation of a diphone inventory. Olive and Spickenagle (1976) attempted to extract the essential features from each diphone by characterizing it in terms of an initial linear prediction pseudo-area function and a linear transition to a final pseudo area function. Diphones obtained from stressed syllables could be used to synthesize new stressed syllables, but the extensive time expansion and time contraction of diphones that is required to satisfy timing rules for stressed and unstressed syllables of English sentences have been a problem. The expected large gain in naturalness that one might expect from utilization of pieces derived from natural speech has not been realized due to compromises that are necessary, such as smoothing at diphone boundaries, changing the duration of the diphones, and imposing a fundamental frequency contour different from that originally recorded (example 22 of the Appendix). At this time, the naturalness of text-to-speech systems based on linear prediction diphones is not significantly better or worse than formant synthesis by rule, in my opinion, although the two types of systems seem to have a different set of perceived deficiencies in naturalness. Diphones must all be recorded by a speaker who can control (hold constant) voice quality so that there aren't sudden changes in the source spectrum in the middle of syllables. But this also means that there is no simple way to change voice quality over a sentence as a function of syllable stress and position within a sentence, leading to a somewhat stereotyped voice quality. The buzziness inherent in LPC also degrades perceived voice quality. On the other hand, a flexible formant synthesizer may permit manipulation of the voicing source characteristics over a sentence, but we do not yet know the rules to do this in an optimal way.

The intelligibility of carefully chosen diphones can be quite high, especially with modern methods, such as the use of multipulse linear prediction (Atal and Remde, 1982) to more accurately characterize noise bursts and other onsets. A third generation of the Olive diphone concatenation scheme is used in an experimental AT&T Bell Laboratories text-to-speech system (Olive and Liberman, 1985) (example 34 of the Appendix). An earlier version of this Bell Laboratories system has been demonstrated for several years at the Epcot Center of Walt Disney World. Conversant Systems, a wholly owned subsidiary of AT&T, has indicated plans to offer for sale a version of this system, although no date has been set for its availability.

A closely related alternative to the diphone is the demisyllable (Fujimura and Lovins, 1978), i.e., half of a syllable. The inventory of half-syllables in English is about 1000 if one is clever about the treatment of certain postvocalic clusters (treating morphemic plural and past consonant sequences such as " — s" and " — t" as separable units, as suggested by Fujimura and Lovins). The advantage of the demisyllable is that highly coarticulated syllable-internal consonant clusters are treated as units, while the disadvantage is that coarticulation across syllables is not treated very well. A synthesis-by-rule program based on demisyllables has been demonstrated by Browman (1980) (example 23 of the Appendix). Perhaps the best choice among concatenation models is a hybrid diphone approach that uses consonant clusters as units when necessary to model the acoustic manifestations of consonant sequences in a satisfactory way (Olive and Liberman, 1979).

In summary, efforts to develop methods for synthesizing phonetic segments to make up arbitrary sentences have proceeded along three lines: creation of (1) heuristic rules for controlling formant synthesizers, (2) "natural" rules for controlling articulatory models, and (3) methods for concatenating pieces of lpc-encoded real speech. The inherent attraction of articulatory solutions must be tempered by practical considerations of computational cost and lack of data upon which to develop rules. The choice between rule systems for formant synthesizers and concatenation strategies may ultimately depend on limits to the flexibility and naturalness of concatenation schemes involving encoded natural speech, but the best current lpc-based systems are quite competitive with the best formant-based rule programs.

D. Prosody and sentence-level phonetic recoding

A sentence cannot be synthesized by simply stringing together a sequence of phonemes or words. It is very important to get the timing, intonation, and allophonic detail correct in order that a sentence sound intelligible and moderately natural, Fig. 24. Prosodic details also help the listener segment the acoustic stream into words and phrases (Nakatani and Schafer, 1978; Svensson, 1974; Streeter, 1978). The following three sections take up these topics in detail.

A pure tone can be characterized in physical terms by its intensity, duration, and fundamental frequency. These induce the sensations of loudness, length, and pitch, respectively. In speech, it is the change over time in these prosodic parameters of intensity, duration, and f_0 that carry linguisti-

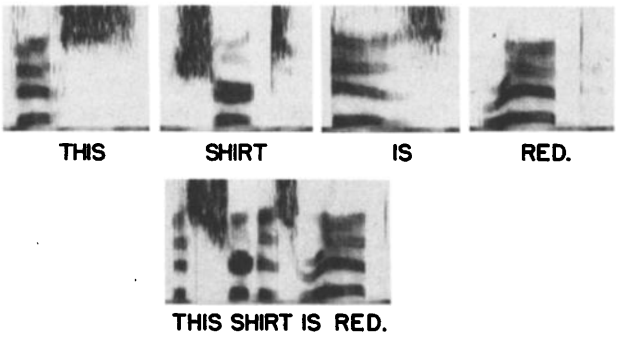

FIG. 24. Broadband spectrograms indicating that a sentence is very different from a concatenated string of words recorded in isolation. Words in sentence context are generally much shorter in duration, are subject to coarticulation at word boundaries, and undergo phonetic recoding—for example, the /t/ in "shirt" has become a flap.

cally significant prosodic information, as summarized in Table I.

Segmental factors that can influence stress judgments include vowel reduction (Fry, 1965) and associated phonological recoding/simplification phenomena. Thus, for example, in the word "photograph," the second vowel is reduced to a short-duration mid schwa vowel [ə], and the /t/ is flapped (compare with "photography").

1. Intensity rules

The intensity pattern of speech tends to set off individual syllables because vowels are usually more intense than consonants. Stressed syllables, which are perceived to be louder than unstressed syllables, may be more intense by a few dB, but intensity *per se* is not a very effective perceptual cue to stress (Fry, 1958), due in part to the confounding variations in syllable intensity associated with vowel height, f_0, laryngeal state, and other factors.

In a formant synthesizer, as in speech, the intensity of a voiced sound automatically goes up in proportion to f_0. Thus one can achieve a degree of stress-related intensity increase by rules that only manipulate f_0. Experience suggests that including a specific rule to increase stressed vowel intensity produces artificially strong stressed vowels.

At a phrase level, it appears that syllables at the end of an utterance can become weaker in intensity, especially if unstressed. However, it is not clear that this is simply an effect of reduced source intensity; usually the glottal waveform becomes more breathy as well, with a strong fundamental component and weaker high-frequency harmonics (Bickley, 1982).

If prosody is to include these source modifications, as it probably should in order to account for natural changes to voice quality over utterances, then we will need new descriptors and new data to quantify the perceptually important effects. At the very least, a new prosodic dimension is required to characterize a continuum of voice qualities from breathy through normal to creaky (Ladefoged, 1973; Catford, 1977). Other possible dimensions might be related to the stability of the vibration pattern (susceptibility to aperiodicities).

TABLE I. Physical and subjective components of sentence prosody.

Physical quantity	Nearest subjective attributes
Intensity pattern	syllabic structure vocal effort, stress
Duration pattern	speaking rate, rhythm, stress, emphasis, syntactic structure
f_0 pattern	intonation, stress, emphasis, gender, vocal tract length psychological state, attitude

2. Duration rules

Aspects of speech timing are specified and modified by information coming from many different representational levels during speech production. Psychological and semantic variables influence the average speaking rate and determine durational increments due to emphasis or contrastive stress. The syntactic structure of the sentence to be produced determines the locations of prosodic boundaries at which segments are longer in duration. The lexicon and/or stress rules determine which consonants and vowels of a word are stressed and hence longer in duration than unstressed and reduced vowels. The phonological component of the speaking process selects appropriate allophones for the abstract phonemes of lexical items, and executes a set of rules that modify the allophone durations according to phonetic context. These effects have been examined in review papers by Lehiste (1970) and by Klatt (1976a).

As an example of the kinds of rules needed to predict segment durations in sentences, consider the model proposed by Klatt (1979a). The model assumes that (1) each phonetic segment type has an inherent duration that is specified as one of its distinctive properties,[6] (2) each rule tries to effect a percentage increase or decrease in the duration of the segment, but (3) segments cannot be compressed shorter than a certain minimum duration (Klatt, 1973b). The model is summarized by the formula:

$$\mathrm{DUR} = \mathrm{MINDUR} + \frac{(\mathrm{INHDUR} - \mathrm{MINDUR}) \times \mathrm{PRCNT}}{100}, \quad (2)$$

where INHDUR is the inherent duration of a segment in ms, MINDUR is the minimum duration of a segment if stressed, and PRCNT is the percentage shortening determined by applying rules described in Table II.

Segmental duration is one of the cues that (1) helps distinguish between segments (e.g., short /ɛ/ versus long /æ/, or short /z/ versus long /s/), (2) determines features of neighboring segments (e.g., the voicing feature of postvocalic obstruents is cued in part by vowel duration—[æ: z] versus [æ s]), (3) distinguishes between stressed and unstressed syllables, (4) signals phrase and clause boundaries, and (5) helps indicate the presence or absence of emphasis. Perceptual disentanglement of these effects is difficult (Klatt, 1982b). In fact, one of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena.

Other durational rule systems exist for English (Mattingly, 1968; Barnwell, 1971; Coker *et al.*, 1973; Umeda, 1975, 1977). The rules contained in these systems are similar (not surprisingly), but there are too many ways to describe interacting phenomena, so that, e.g., Gaitenby *et al.* (1972) and Coker *et al.* (1973) rely heavily on multiple stress levels conditioned by syntactic category (verbs have less stress than nouns) and conditioned by word frequency (common words and words that are repeated in a discourse are reduced in stress). Other authors postulate rules related to rhythm and isochronous principles (Lehiste, 1977). Neither of these

TABLE II. Duration rules proposed by Klatt (1979a).

1. PAUSE INSERTION RULE: Insert a brief pause before each sentence-internal main clause and at other boundaries delimited by an orthographic comma (Goldman-Eisler, 1968; Cooper *et al.*, 1978).
2. CLAUSE-FINAL LENGTHENING: The vowel or syllabic consonant in the syllable just before a pause is lengthened (Gaitenby, 1965). Any consonants in the rhyme (between this vowel and the pause) are also lengthened (Oller, 1973; Klatt, 1975a).
3. PHRASE-FINAL LENGTHENING: Syllabic segments (vowels and syllabic consonants) are lengthened if in a phrase-final syllable (Klatt, 1975a). Durational increases at the noun-phrase/verb-phrase boundary are more likely in complex noun phrase or when subject–verb–object order is violated; durational changes are much less likely for pronouns (Harris *et al.*, 1981). The lengthening is perceptually important (Lehiste *et al.*, 1976; Umeda and Quinn, 1981).
4. NON-WORLD-FINAL SHORTENING: Syllabic segments are shortened slightly if not in a word-final syllable (Oller, 1973). [This rule is disputed by Umeda (1975).]
5. POLYSYLLABIC SHORTENING: Syllabic segments in a polysyllabic word are shortened slightly (Lehiste, 1975a). [This rule is also disputed by Umeda (1975).]
6. NON-INITIAL-CONSONANT SHORTENING: Consonants in non-word-initial position are shortened (Klatt, 1974; Umeda, 1977).
7. UNSTRESSED SHORTENING: Unstressed segments are shorter and more compressible than stressed segments (Fry, 1958; Umeda, 1975, 1977; Lehiste, 1975a).
8. LENGTHENING FOR EMPHASIS: An emphasized vowel is significantly lengthened (Bolinger, 1972; Umeda, 1975).
9. POSTVOCALIC CONTEXT OF VOWELS: The influence of a postvocalic consonant (in the same word) on the duration of a vowel is such as to shorten the vowel if the consonant is voiceless (House and Fairbanks, 1953; Peterson and Lehiste, 1960). The effects are greatest at phrase and clause boundaries (Klatt, 1975a).
10. SHORTENING IN CLUSTERS: Segments are shortened in consonant–consonant sequences (disregarding word boundaries, but not across phrase boundaries) (Klatt, 1973a; Haggard, 1973).
11. LENGTHENING DUE TO PLOSIVE ASPIRATION: A stressed vowel or sonorant preceded by a voiceless plosive is lengthened (Peterson and Lehiste, 1960).

kinds of rules is incorporated explicitly in the Klatt system, but partial isochrony is achieved through rules that shorten unstressed syllables and consonant clusters (Carlson *et al.*, 1979). The Klatt rules capture durational differences between nouns and verbs by phrase-final lengthening and destressing of common verbs. An emphasis symbol is provided to capture word frequency and discourse expectancy effects in a binary fashion. These alternative mechanisms for mimicking observed tendencies in durational data make it nearly impossible to determine which rule system has a basis most similar to psychological processes.

3. Fundamental frequency rules

Many phenomenological observations have been collected about pitch motions in English sentences, and hypotheses have been generated concerning their relations to linguistic constructs known as intonation and stress. The intonation pattern is defined to be the pitch pattern over time that, for example, distinguishes statement from question or imperative, and that marks the continuation rise between clauses for an utterance of more than one clause. The stress pattern on syllables can distinguish words such as “ ’insert” from “ins’ert” even though the two words have identical segmental phonemes. Linguists originally believed that there was a fairly direct correspondence between intonation and pitch, while levels of stress were manifested by changes in vocal intensity and syllable duration. Now we know that f_0 *changes* affect stress judgements significantly (Fry, 1958; Nakatani and Schafer, 1978), and that a rise in f_0 or a fall in f_0 can indicate a stressed syllable. The f_0 pattern plays a complex role in encoding information for the listener because it not only conveys information about syntactic structure and stress patterns, but it also helps indicate speaker gender, head size, psychological state, and attitude toward what is being spoken. This section reviews briefly some of what is known about this encoding.

Pike (1945) believed that English is like a tone language in that four different degrees of stress corresponded to different pitch levels. However, it has been shown that a given stress level is manifested as a higher pitch at the beginning of a sentence than near the end (Lieberman, 1967), so absolute f_0 cannot be the relevant cue to the level of a tone. Lieberman also demonstrated that (simulated) emotional states changed f_0 patterns in ways that made it impossible for linguists to assign stress levels to syllables in a consistent way when listening to read sentences. Thus emotions and attitudes are also conveyed to some extent by f_0 patterns (for sample data, see Uldall, 1960; O'Shaughnessy and Allen, 1983). Instrumental analyses also indicated that segmental identity could perturb the f_0 value (House and Fairbanks, 1953), and that there were large differences across speakers depending primarily on larynx size. On average, female speakers use f_0 values about 1.7 times male values (Peterson and Barney, 1952), plus perhaps a slightly more lively set of dynamic changes (higher peaks and lower troughs) than simple scaling would imply.

Bolinger (1972) notes the frequent use of contrastive stress or emphasis in expressive reading. To the extent that locations for emphasis can be determined for text, the emphasis can be manifested acoustically by increasing the duration of the emphasized word, increasing the pitch rise that ordinarily accompanies its primary-stressed syllable, and decreasing the size of all other pitch rises in the remainder of the sentence (Cooper and Sorenson, 1981).

O'Shaughnessy (1979) and O'Shaughnessy and Allen (1983) examined f_0 contours for syntactically complex sentences, and for sentences involving modals. They observed that modal auxiliaries, negatives, quantifiers, and sentential adverbs tend to be emphasized (local f_0 increase) when present in read sentences. The authors interpret these results in terms of the speaker's feeling toward the proposition tending to dominate over the actual content of the proposition (Halliday, 1970).

The strength of an f_0 gesture depends on semantic factors that extend over more than one sentence (Coker *et al.*, 1973). A repeated word is reduced in f_0 gesture, and the reduction is due to semantic recurrence rather than to reappearance of exactly the same item (Vanderslice, 1968). In addition, the structure of discourse seems to cause readers to start with a higher f_0 at the beginning of a paragraph (Lehiste, 1975b).

In addition to the rule-governed changes to fundamental frequency over a sentence, there are also local perturbations due to aspects of segmental articulation. The f_0 contour is higher near a voiceless consonant than near a voiced consonant, and is higher on a high vowel (House and Fairbanks, 1953; Peterson and Barney, 1952), although this latter effect may be reduced in sentence contexts (Umeda, 1981).

For synthesis by rule, what is needed is a theory that can predict when f_0 will rise or fall, and what levels it will reach on individual stressed syllables of a sentence as a function of syntactic structure, stress pattern, and semantic/performance variables (if known) such as the location of the most important word in the sentence, or the speaker's attitude toward what is being said. Such theories are still in their infancy, and many alternative formulations exist, but fortunately several are complete enough to serve as models for a text-to-speech algorithm. One type of theory is based on the view that f_0 moves (sluggishly) from target to target tone (Pike, 1945). Another class of theories includes commands to raise and lower f_0 at certain times, emphasizing the motion over the actual target achieved (Bolinger, 1951), see also Ladd (1983).

The first algorithm for determination of a fundamental frequency contour was programmed by Mattingly (1966) and incorporated in the phonemic synthesis-by-rule program of Holmes *et al.* (1964). In the British tradition of Armstrong and Ward (1931), which separates intonation and stress, Mattingly's rules recognized three intonational "*tunes*" that could be placed on the last prominent syllable of a clause. The tunes, shown at the top in Fig. 25, are "*falling*," "*rising*," and "*fall–rise*"—corresponding to statement end, question end, and continuation rise. Other prominent syllables of a sentence (typically the stressed syllable in semantically important content words) could be marked by the user; in which case these received a local increase in f_0. Unstressed syllables were generally lower in pitch because they were not assigned a target.

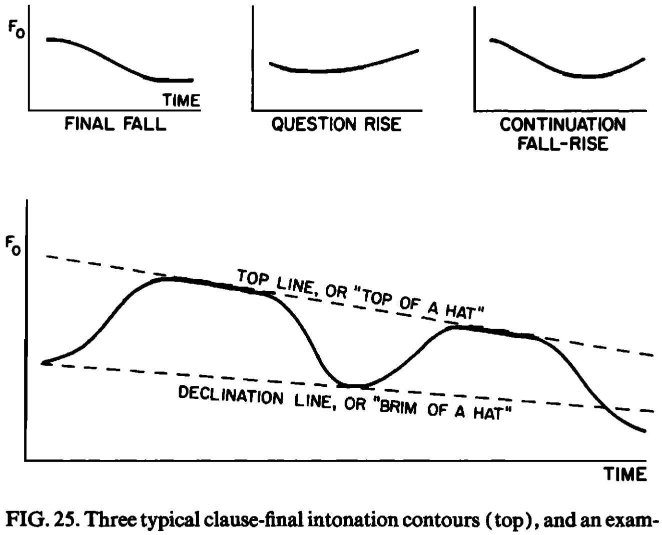
F0
TIME
FINAL FALL
QUESTION RISE
CONTINUATION FALL-RISE
F0
TOP LINE, OR "TOP OF A HAT"
DECLINATION LINE, OR "BRIM OF A HAT"
TIME

FIG. 25. Three typical clause-final intonation contours (top), and an example of a fundamental frequency "hat pattern" of rises and falls between the brim and top of a hat for a two-clause sentence (bottom).

These rules were intended to mimic intonation patterns of British English; an American version was published later by Mattingly (1968). In this rule system, the tendency for f_0 to start high and fall gradually throughout a sentence (declination) was reduced for American English, and the prominent/nonprominent opposition was elaborated by distinguishing three stress levels (primary, secondary, and unstressed).[7] The influence of consonants on f_0 (Lehiste and Peterson, 1961) was approximated by causing the f_0 to start higher at the onset of a stressed syllable if it began with a voiceless consonant.

A similar view of intonation was described in quite different terminology by 't Hart and Cohen (1973). In the spirit of Bolinger (1951), they defined the intonational "*hat pattern*," see bottom portion of Fig. 25, as the tendency for intonation to rise on the first stressed syllable of a phrase, and remain high until the final stressed syllable where there is either a dramatic fall or a fall–rise depending on whether more material is to be spoken. The idea of intonational phrases is similar to the idea of the breath group advocated earlier by Lieberman (1967). Translation of these ideas to rules for English was performed by Maeda (1974), who also postulated stress-related local rises above the phrasal hat top whose magnitudes depended on phrasal position—the size of pitch gestures tending to be reduced over the course of a phrase.

The Maeda rules form the basis for the f_0 gestures produced by Klattalk. The detailed implementation is based on an idea of Öhman (1967). He proposed that intonation contours can be modeled in terms of impulses and step commands fed to a linear smoothing filter. This type of model has been applied to Japanese intonational synthesis by Fujisaki and Nagashima (1969), who were able to match natural intonation contours with remarkable fidelity. An example of the step and impulsive commands for a sentence generated by Klattalk rules is shown in Fig. 26.

The timing of the fundamental frequency rises and falls with respect to the locations of stressed vowels can have a fairly large perceptual effect. For example, gradual rises extending over the full vowel duration are heard as similar to continuation rises—indicative of material prior to the most prominent or nuclear syllable of the utterance.

The most detailed current model of f_0 generation for American English (Pierrehumbert, 1981; Anderson *et al.*, 1984) takes a somewhat different approach to the problem, and posits two f_0 target tones at an abstract level—H (high) and L (low). Each stressed syllable of a sentence is assigned a sequence of zero or one such tones according to syntax, discourse importance, and rhythmic position. In addition, there are two extra tones at the end of a phrase, one occurring between the last accent and the end, and the other occurring right at the end. These permit various forms of terminal falls and rises to be constructed. The assignment of f_0 targets and smooth transitions between targets is a complex function of a reference f_0 declination line (Öhman, 1967; Peck, 1969) and a time-varying pitch range (Cohen and 't

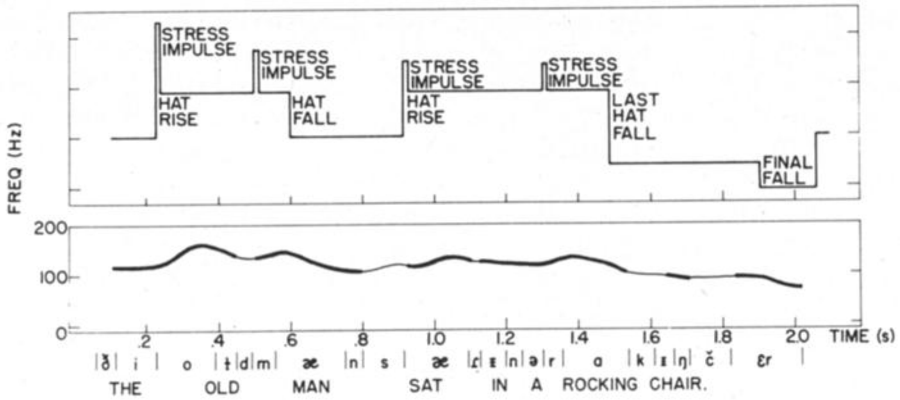

FIG. 26. The Klattalk f_0 contour shown at the bottom was generated by sending a sum of the various step and "impulsive" commands shown above through a low-pass smoothing filter. Additional small step commands associated with tongue height and glottal state, and a gradual declination line also served as input to the low-pass filter.

Hart, 1967). The model can deal with a wide range of observed intonational patterns, but many of the patterns could only be predicted from text if one were a mind reader (Bolinger, 1972). A stripped-down version of the model is used in the Bell Laboratories text-to-speech system described earlier. Demonstrations of the system (example 34 of the Appendix) use input text where adjective–noun and compound stress patterns are hand corrected if necessary, because getting this aspect of prosody correct is both difficult and perceptually quite important.

It can be frustrating to work with rule systems for generation of f_0 and duration patterns for sentences in a text-to-speech context because one depends on sentence analysis routines to determine aspects of syntactic structure or semantic importance, and these routines are often wrong. When a text-to-speech system makes a phonemic pronunciation error, the user may be able to override the text-to-phoneme process by re-specifying the word phonemically. Fortunately, in some systems, the same type of user correction capabilities exists for prosodic errors. For example, DECtalk permits syntactic symbols to be placed in the orthographic or phonemic transcription. If this does not lead to a better prosodic reading, the device will accept durations, specified in ms, for any input phonetic segment (Conroy *et al.*, 1986). A hand-drawn fundamental frequency contour can also be specified by straight-line interpolation between f_0 targets specified at the end of each phonetic segment. Fairly natural prosody can be achieved by the painstaking copying of a recorded utterance using these facilities.

4. Allophone selection

We have assumed that words are lexically represented by phonemes and stress symbols. Allophone selection is then an important aspect of the sentence generation process. For example, the word "city" might appear in a pronouncing dictionary as/s'ɪti/, i.e., with a medial /t/ phoneme, but the word is almost always pronounced with a flap variant [ɾ] of the /t/, see Fig. 27. It might appear possible to obviate the need for a flapping rule by simply representing "city" with a flap in the first place. However, a flap rule is still required in a text-to-speech system in order to turn the fully released [t] of "bait" into a flap in a phrase such as "bait a hook." Slightly oversimplifying, a /t/ is flapped in American English between two sonorants if the second is unstressed. At least for those cases where a phoneme can take on different allophones depending on the context of the word, a set of allophone selection rules is unavoidable. Cross-word-boundary phonological recoding is significant in English, as we will see.

Part of the problem of speaking naturally concerns the phonetic form of function words. Words such as "for," "to," "him" often take on the reduced forms [fɚ], [tə], and [ɨm] (Heffner, 1969), but not in all phonetic environments. For example, in Klattalk, "for" is not reduced if the next segment is a vowel or silence. If these words are never reduced, the speech sounds stilted (something like that of a bad actor trying to articulate carefully), while over-application of rules for reducing function words may lead to misperceptions as to the number of syllables in an utterance.

While a phoneme inventory for English can be specified with little debate, selection of an appropriate inventory of allophonic symbols involves many conflicting criteria and tradeoffs. The clearest cases are those where a phoneme is

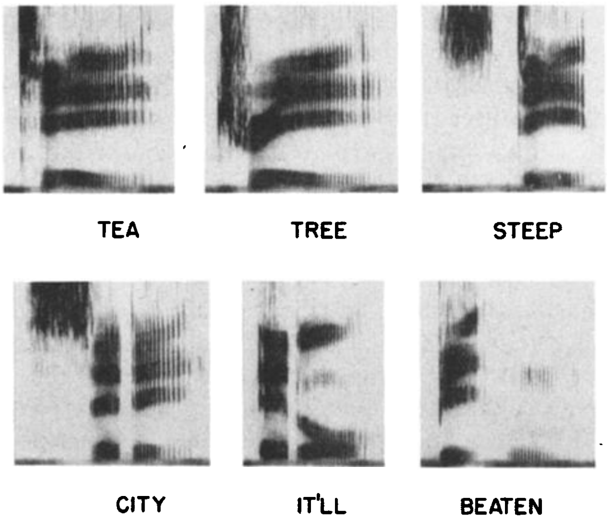

FIG. 27. Broadband spectrograms illustrating allophonic changes to /t/: a normal aspirated /t/ in "tea," affricated in "tree," unaspirated in "steep," flapped in "city," lateral or glottal release in "it'll," and nasal or glottal release in "beaten," after Zue and Laferierre (1979).

replaced by an allophone with distinctly different articulatory/acoustic properties. For example, the phoneme/l/ is realized as a velarized variant following a vowel, while there is normally no velar constriction for word-initial productions of /l/ (Lehiste, 1962).

Less clear are those cases where a small change is the result of a low-level articulatory interaction (Schwartz, 1967), or where many small changes can be made along an articulatory/acoustic dimension such as voice onset time. For example, the time between release of a /t/ and voicing onset is typically about 50 ms, but is systematically about 10 ms longer in a word-initial position, e.g., "tone," than it is in prestressed word-medial positions, e.g., "atone," and VOT is shorter if the following vowel is unstressed (Klatt, 1975b). Should one create a separate symbol for each gradation along the voice-onset-time continuum, or handle these effects as low-level adjustments to the time functions that control the synthesizer? Distinctions between allophone selection rules and parameter adjustment rules are necessarily arbitrary, and of relatively little theoretical import to us.[8] The important thing is to be able to produce the appropriate acoustic changes in the synthetic speech, and to do so in an efficient way.

Some of the rules to be discussed below appear to be articulatory simplifications that allow the speaker to be "lazy" in realizing some unstressed phonetic sequences. While ease of pronunciation may play a role in the development of allophonic variation, a far more important function of these rules is to help mark boundaries, especially word boundaries, in the flow of speech (Lehiste, 1959; Nakatani and Dukes, 1977); Lehiste cites many examples where allophones mark boundaries, the best known of which is the distinction between "night rate" and "nitrate," where a listener can easily tell which sequence was intended by the speaker because of stronger frication/aspiration in the latter case.[9]

Most of the rules discussed below are thus not strictly "sloppy speech" rules, and they are not optional rules. They are needed to make sentences sound fluent and natural. The rules help the listener decide the syllable affiliation of consonants and the degree of stress on a syllable, and thus indirectly constrain locations of potential word boundaries, permitting the listener to parse an utterance into words without pursuing too many alternative interpretations (Church, 1983). Phonotactics, or the specification of permitted phonetic sequences at the beginnings, middles, and ends of words, also can provide word boundary hypotheses for the listener (Lamel and Zue, 1984).

The details of phonological rule application differ for the different dialects of English, as well as for different speaking styles (formal/casual) and speaking rates within a given dialect. This is a serious problem for speech recognition devices (see, e.g., the rule compendium of Cohen and Mercer, 1974), but a text-to-speech system need only select rules appropriate for one acceptable dialect of English, and perhaps make some modifications concerning rule applicability as a function of speaking rate (Bernstein and Baldwin, 1985). In Klattalk, some phonetic simplifications across word boundaries are blocked if a phrase boundary is present. This mechanism is used to produce more formal speech at slow speaking rates simply by placing phrase boundary symbols at more minor phrase breaks when analyzing a text. In the future, it might be interesting to attempt to simulate additional dialects and styles by direct manipulation of phonological rules in these systems.

Some of the allophonic phenomena to be described have been known for a long time, many having appeared in phonetics textbooks at least as far back as the 1930's (Bloomfield, 1933; Hocket, 1955; Heffner, 1969). However, acoustic characterization had to await instrumental study. One of the first and best of the acoustic–phonetic studies was performed by Lehiste (1959, 1964). She noted the following kinds of word boundary indicators:

- The presence of a laryngealized vowel onset usually signals the beginning of a word that starts with a vowel.
- A normally aspirated release of [p,t,k] becomes unaspirated if a preceding [s] is part of the same word ("the spot" versus "this pot").
- Selection between an initial or final allophone /r/ or /l/ intervocalically depends on the location of a word boundary on either side of the consonant.
- A vowel is longer in duration in an open syllable (no word-final consonant), and shorter if followed by a voiceless word-final consonant.
- A word-final [t,d] is flapped or glottalized before a word beginning with a stressed vowel.

Nakatani and O'Connor-Dukes (1979) extended this work, and concluded that the phonetic cues and stress changes are perceptually more powerful cues to word boundary locations than are durational and pitch changes associated with syntactic boundary movements. They used an analysis–resynthesis system to generate stimuli with, e.g., durational characteristics of one phrase and phonetic characteristics of another in order to obtain perceptual judgments of cue strength. Additional phenomena that they noted include:

- Geminate consonants are lengthened with respect to singletons (e.g., the /k/ in "drunk converse" versus "drunken verse").
- Vowels can be deleted and words resyllabified (e.g., when "bakery" becomes a two-syllable word).
- There are restrictions on vowel reduction (e.g., there is reduction in "hard *de*feat" but not in "hardy feet").

In a subsequent study of [l] and [r], Lehiste (1962) noted that the prevocalic "light" allophone of /l/ as in "lead" has a second formant that depends on the following vowel, postvocalic "dark" or velarized /l/ as in "deal" has a lower second formant that is independent of the preceding vowel, and is similar to the syllabic /l/ in "bottle." The initial allophone of /r/ as in "reed" has lower $F1$, $F2$, $F3$ than the postvocalic allophone as in "deer." The syllabic nucleus [ɝ] as in "dirt" has formant targets similar to the postvocalic allophone.

In a study of the allophones of /t,d/ and their distribution, Zue and Laferriere (1979) distinguished:

- within-word prestressed variants as in "return" and "reduce,"

- unstressed (shorter, less aspirated) versions as in "minty," "Mindy," "moulted," and "molded,"
- voiced flaps as in "rater" and "raider,"
- glottalized or nasal released stops, as in "sweeten" and "Sweden,"
- deleted allophones, as sometimes occur in "pentagon."

Klatt (1975b) measured burst durations and voice onset times (VOT) for plosives in consonant clusters as a function of stress/phonetic/structural environments, and proposed a set of quantitative rules to account for the data. As is well known, VOT for /p,t,k/ is longer in clusters with a following sonorant consonant, shorter in a cluster with a preceding /s/, shorter if the syllable is unstressed, longer in word-initial position, and shorter if preceded by a voiced segment of a preceding word. Most of the rules are natural consequences of aerodynamic factors involved in getting the glottis open in order to generate aspiration, and then closed to begin voicing. For example, Umeda and Coker (1974) observed that the duration of aspiration for prevocalic [t] tends to covary with closure duration, and that VOT is shorter for (unstressed) function words like "to" than for content words like "two."

Morpheme structure can be important in determining the acoustic realization of consonants. For example, /p,t,k/ are not strongly aspirated in /sp,st,sk/ clusters, except for the case where there is an obvious morpheme boundary after the /s/, as in, e.g., "discourteous" and "miscalculate" (Davidsen-Nielsen, 1974). The morpheme boundary symbol must be present in the abstract linguistic description for such words if the aspiration feature is to be computed correctly. Otherwise, a principle of assigning the maximum number of prevocalic consonants to a medial stressed vowel, subject to the constraint that the consonants form a legitimate word-initial consonant cluster (Hoard, 1966, 1971), will group the/s/-plosive into a prestressed cluster. This syllabification principle is used in Klattalk, resulting in reduced aspiration for [p,t,k] in words like "discourteous" unless a morpheme boundary is inserted after the /s/.

Prevocalic and postvocalic allophones may differ in acoustic aspects related to the temporal buildup/decay of the sound source. Coker and Umeda (1975) observed that the prevoicing for [b,d,g] is weaker and less rich in higher harmonics in utterance-initial positions due to the more sinusoidal nature of vocal fold vibrations at initiation of voicing. Similarly, [m,n,l] were a few dB weaker in intensity (during the early portion of the consonant) in word-initial positions than in medial and final positions. On the other hand, the noise intensity for [s] was about 3 dB more intense word initially than medially and finally, presumably due to the slightly higher subglottal pressure (or the timing or pressure buildup/decay) ,associated with initial versus utterance-final consonants (Umeda and Coker, 1974).

In a search for sentence-level recoding rules, Oshika *et al.* (1975) noted the palatalization of word-final alveolar consonants if the next word begins with a palatal consonant, as in "did you" [dɪǰu] and "this shoe" [ðɪššu]. Zue and Shattuck-Hufnagel (1979) found the effect to be asymmetrical, applying to the [s] in "this shoe" but not to either the [s] or the [š] of "wish some."

Broad and Fertig (1970) examined a collection of about 150 different C_i-ɪᵊ-C_f nonsense words spoken by a single trained speaker. They measured formant values at ten equally spaced locations throughout each syllable, and then performed averaging over time and tokens to obtain formant values associated with the vowel. Next, they measured an average formant transition for each initial consonant, C_i, averaged over all possible final consonants, C_f. They represented this transition as a difference between the measured trajectory and the average formant position for [ɪᵊ]. They observed that formant transitions associated with plosives were generally restricted to about half the vowel duration, as shown at the top in Fig. 28, but sonorant consonants often affected the entire vowel. They tried to determine whether average formant transitions for each initial C and each final C were sufficiently regular that one could predict in detail the whole formant pattern for each individual syllable from a sum of the average [ɪᵊ] trajectory and the superimposed incremental trajectories for the initial and final consonants, as

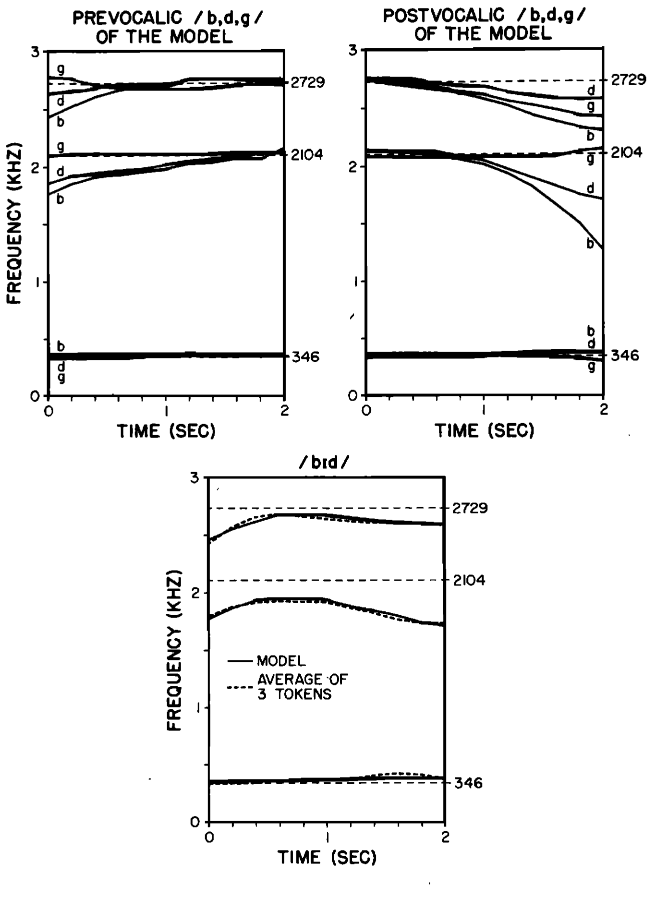

FIG. 28. Approximation to the formant transitions for the CVC syllable [bɪd] (bottom panel) was derived in terms of a model (top panels) that takes into account the average vowel formant positions (dashed lines) and incremental perturbations due to each consonant (solid lines), after Broad and Fertig (1970). At least for the vowel /ɪ/, the data suggest that formant motions can be additively decomposed into (1) an underlying vowel target, (2) a transition associated with the prevocalic consonant, and (3) a transition associated with the postvocalic consonant.