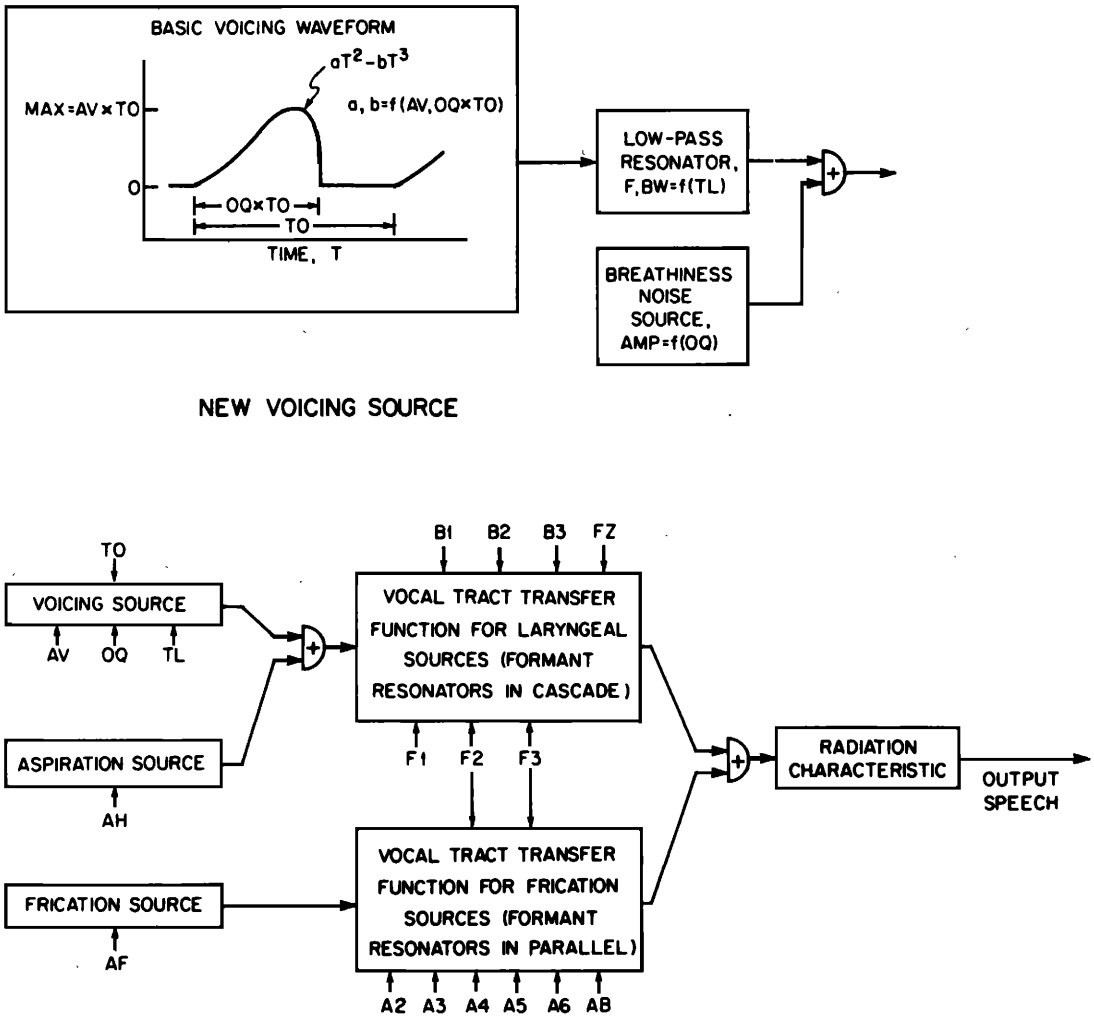

FIG. 12. Block diagram of the Klattalk synthesizer in which a new voicing algorithm (top) has been added to the synthesizer (bottom) that was described in Klatt (1980). Nineteen variable control parameters are identified, including the new voicing source parameters OQ (open quotient) and TL (spectral tilt). Other synthesizer constants that are not shown, such as the frequencies of the fixed fourth and fifth formant resonators, can be reset by the user by modifying a set of speaker-defining constants.

lar devices suggests that either something is still missing from the voicing source models, or that we do not yet know how to control them properly.

A number of recent glottal waveform models produce source spectra that include zeros (see Fujisaki, 1986 for a review). Flanagan (1972, pp. 232–245) describes the expected locations of voicing source spectral zeros as a function of various assumptions about the nature of the glottal volume velocity waveform. Many different types of waveshapes imply the existence of zeros; the only requirement is that there be well-defined open and closing times. If a source spectral zero is near in frequency to a formant, the formant will be reduced in amplitude or even completely obliterated. Source spectral zeros are present in the glottal waveform models of Fant *et al.* (1985) and in Klattalk, but the depth of the spectral notches is only a few decibels. Flanagan shows that the frequency locations and depth of spectral notches induced by source zeros depend on relatively small changes to critical aspects of the source waveform, such as symmetry. It may be that the dull, lifeless quality of synthetic voices is due in part to the absence of small period-to-period changes to the zero pattern. Holmes (1973) was able to synthesize a nearly perfect imitation of a male voice without resorting to this level of detail in modeling the source, but he may have mimicked the most important effects of source changes by ensuring that the amplitudes of individual formant spectral peaks followed changes observed in the natural utterance.

Naturalness is a particular problem when trying to synthesize a convincing imitation of a female voice (Carrell, 1984). Simple scaling procedures [formants multiplied by a factor of 1.15 (Peterson and Barney, 1952), fundamental frequency by a factor of 1.7, glottal open quotient slightly greater than for a male voice] do not result in a particularly female voice quality (example 9 of the Appendix). The glottal source model is not quite right; nonuniform formant scaling appears to be required (Fant, 1975), and it may also be

that men and women adopt certain speaking strategies and dialectal differences to signal their gender (Kahn, 1975; Labov, 1986).

Based on a detailed spectral analysis of a single female speaker having a pleasant voice quality (Klatt, 1986b), I have begun efforts to synthesize a copy of some of her utterances using the flexibility of the new Klattalk voicing source. The analysis revealed the presence of considerable random breathiness noise at frequencies above 2 kHz over portions of many utterances (a possibility noted earlier by Fujimura, 1968), and considerable variation in both the general tilt of the harmonic spectrum and the strength of the fundamental component. When these factors are modeled in the synthesis, by varying the open quotient, spectral tilt, and breathiness noise amplitude parameters of the Klattalk voicing source, Fig. 12, very good approximations to this voice are achieved for isolated vowels. Success was achieved even though I used a cascade synthesizer rather than the parallel configuration advocated by Holmes, and therefore did not have direct control over each formant amplitude. Also, for at least this one voice, the source spectral zeros seemed to be well matched in location and depth with respect to observed natural spectral dips, even though only the open quotient parameter was available as a means of adjusting the frequency positions of the zeros.

In order to see if the preliminary success with isolated vowels could be generalized to more complex speech materials, the next step taken was to analyze a set of reiterant sentences that were spoken by replacing all of the intended syllables by [ʔV] or [hV], where [V] was one of six English vowels. Utterances involving a glottal stop were considerably easier to model (example 10 of the Appendix). The vowel spectra generally conformed to the simplified acoustic theory implicit in the synthesizer. However, in the [hV] materials, many of the voiced intervals revealed additional formant peaks and other harmonic amplitude discrepancies, presumably related to acoustic coupling with tracheal resonances when the glottis is partially open. An example is shown in Fig. 13. My best synthesis efforts that did not contain these irregularities were judged to be less human and less like the speaker than in the case of the glottal stop syllables.

These results suggest that spectral details in the mid and low frequencies can be of considerable importance to speaker identity and to naturalness judgments, especially in a female voice, where harmonics are widely spaced and more easily resolved by the auditory system. At this point, it is hard to decide how best to augment the synthesizer in order to model the sudden appearance of additional formants and zeros in breathy vowels. For example, would one additional pole-zero pair be sufficient to approximate the primary perceptual effects of tracheal interactions? Also needed are data upon which to base rules for positioning additional resonance peaks and dips as a function of presumed glottal state and vocal tract shape (it is tough enough estimating formant frequencies in high-pitched voices—to require the simultaneous detection of an unknown number of additional pole-zero pairs as well as specification of glottal source parameters may be asking too much). Nevertheless, a preliminary

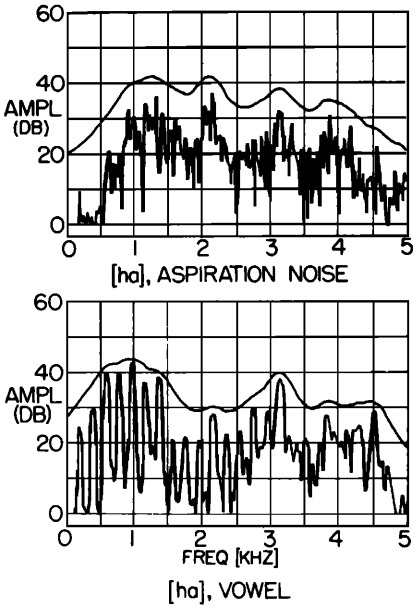

FIG. 13. The magnitude spectrum of 50 ms of aspiration noise in the syllable [ha] (top) reveals a strong subglottal formant at 2150 Hz, between $F2 = 1300$ Hz and $F3 = 3200$ Hz. Additional subglottal resonances are faintly evident at about 600 and 1600 Hz. The effect of the subglottal system on the vowel spectrum, measured about 40 ms after voicing onset (bottom) is to create a spurious peak at 2150 Hz, and to modify harmonic amplitudes in the $F1$–$F2$ region so as to make it difficult to tell whether two or three formants are present between 600 and 1400 Hz. These changes are typical of normal breathy vowels of women (Klatt, 1986b).

attempt to analyze and synthesize a full sentence using a synthesizer configuration augmented by an extra tracheal pole-zero pair (first part of example 10 of the Appendix) has met with some success.

An alternative solution to the problem of producing a natural female voice quality by a formant synthesizer might be to employ articulatory models of the trachea, vocal folds, and vocal tract, as well as their interactions, in a sophisticated articulatory synthesizer. Thus we now turn to efforts to produce speech by direct simulation of the mechanisms involved in speech generation.

4. Articulatory models

The transfer function of the vocal tract can be modeled by formant resonators, as above, or by a direct transmission line analog of the distribution of incremental pressures and volume velocities in a tube shaped like the vocal tract. In an articulatory model the tube corresponding to the vocal tract is usually divided into many small sections, and each section is approximated by an electrical transmission line analog (Dunn, 1950; Stevens *et al.*, 1953). The equations are summarized in Flanagan (1972).

These first electronic models were static and required the hand adjustment of a variable inductor in each section. The possibility of dynamic control was added to the M.I.T. model of Stevens *et al.* (1953) by Rosen (1958). The electronic circuits, shown in Fig. 14, included a buzz source for voicing, and the ability to inject a noise source at the location of a constriction in the vocal tract. Hecker (1962) added a side-branch to approximate the nasal tract. In 1961 at the fall meeting of the Acoustical Society of America, Kenneth Stevens and Arthur House demonstrated that such models

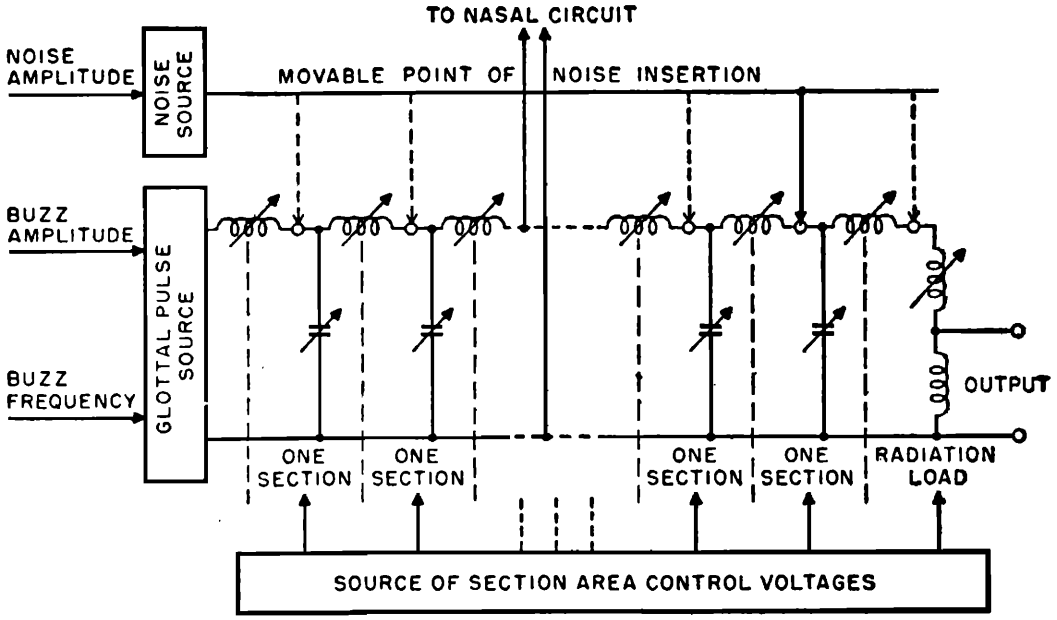

FIG. 14. The DAVO (Dynamic Analog of the VOcal tract) synthesizer, consisting of a ladder network of inductors and capacitors, each section of which mimics the properties of a short section of the vocal or nasal passages, after Rosen (1958). Inductance and capacitance values are determined by the area of vocal tract at that point.

were capable of synthesizing intelligible speech (example 11 of the Appendix).

Modern improved simulations of an articulatory vocal tract have been concerned with the incorporation of frequency-dependent loss terms, provision for cavity wall motion at low frequencies, and better modeling of the time-varying termination impedance at the glottis (Flanagan *et al.*, 1975; Liljencrants, 1985).

The first articulatory synthesizers used a glottal waveform consisting of a sawtooth current source. The voicing source has traditionally been described as a current source because the volume velocity waveform was said to depend very little on the shape or impedance of the vocal tract, at least for vowels (Fant, 1960; Flanagan, 1972). Efforts to improve upon this source model initially focused on obtaining a better approximation to the vibration pattern and resulting volume velocity waveform, while more recently, interactions between source and vocal tract have become of primary concern (Fant *et al.*, 1985).

The first mechanical model of the vibrating vocal folds was a single mass-spring-damping system (Flanagan and Landgraf, 1968). Waveforms generated by this model bore many similarities to physiological data, but the conditions under which the system would vibrate were somewhat restricted. An important aspect of natural vibrations appears to be the out-of-phase motions of the upper and lower surface of the folds (Ishizaka and Matsudaira, 1968; Stevens, 1977; Broad, 1979), and the vertical component to the vibration pattern of the folds (Baer, 1981). The first-order aspects of these phenomena have been captured by two-mass models of each fold, in which the upper and lower surfaces of the folds are simulated by separate masses coupled by a spring (Ishizaka and Flanagan, 1972). The sound generation capabilities of such a model (coupled to a digital simulation of a transmission line analog of the vocal tract) were demonstrated by Flanagan *et al.* (1975) (example 12 of the Appendix).

Another approach to the modeling of the vocal fold vibration behavior has been to create a three-dimensional structure consisting of a large number of coupled masses (Titze, 1974; Allen and Strong, 1985). More complex vibration modes are seen in this type of model, and it may be possible to mimic certain pathologies. However, in all of the physiological models, no entirely satisfactory solution has been proposed for simulating what happens when the vocal folds slam together at the midline and deform in some way to absorb the energy of the impact. Until such phenomena are included, it is difficult to predict when the folds will open or to predict their initial opening velocity (Stevens, 1987).

The resonance structure of the vocal tract results in standing pressure waves that can have an effect on the pressure distribution at the glottis, and hence the vibration pattern and airflow waveform from the voicing source (Fant, 1982; Fant *et al.*, 1985). Similarly, the opening and closing of the glottis provide a time-varying termination impedance that affects the formant frequencies and bandwidths of the vocal tract transfer function (Holmes, 1973; Fant and Ananthapadmanabha, 1982). While these effects are not large, they may be of some importance in simulating natural voice qualities by providing period-to-period variability to the glottal waveform for the first few periods at the onset of voicing, as well as causing pitch-synchronous changes to the first formant frequency and bandwidth over a pitch period. Liljencrants (1985) has programmed a detailed articulatory model to simulate these effects, with the result that the synthesis of a steady vowel sounds quite natural.

The precise acoustic aspects of a complex articulatory model that might account for naturalness (spectral zero movements, glottal waveform changes from period to period, pitch-synchronous formant motions, natural voiceless–voiced–voiceless transitions, etc.) are not known at this time. Also, the considerably greater computational cost of articulatory synthesis precludes the use of these models in practical systems at the present time.

5. Automatic analysis/resynthesis of natural waveforms

Waveform encoding techniques will not be considered in this review (for example, see Lee and Lochovsky, 1983), but perhaps we should note the Texas Instruments "Speak'n Spell" toy (Wiggins, 1980), which used linear prediction encoding (Itakura and Saito, 1968; Atal and Hanauer, 1971; Markel, 1972; Makhoul, 1973) to store and play back a set of words at a storage cost of about 1000 bits/s of speech (example 13 of the Appendix). This inexpensive device has had a major impact on the technology of presenting "canned" messages to the public. Linear prediction representations of speech waveforms are based on the idea that, at least in the absence of source excitation, the next sample of a speech waveform can be estimated from a weighted sum of 10 or so previous waveform samples, the weights being the linear predictor coefficients. If the source waveform can be found by other means, and if predictor coefficients are updated every 10 ms or so on the basis of analysis of a speech waveform, reasonably good approximations to the original waveform can be derived from this kind of low bit rate representation.

In a text-to-speech application, it is necessary to employ an analysis/resynthesis procedure that will allow the natural speech samples to be modified in fundamental frequency, amplitude, and duration, as well as perhaps performing some sort of parameter smoothing at boundaries between waveform pieces. Linear prediction analysis of speech appears to be an excellent representation for these purposes (Olive and Spickenagle, 1976). It is even possible to reconstruct a waveform that is perceptually nearly indistinguishable from the original if multipulse excitation (Atal and Remde, 1982) is used to correct some of the errors that occur when the vocal tract is not all-pole and when the glottal source waveform is not like an impulse train (example 14 of the Appendix).

However, a problem with this approach arises when going from duplicating a natural utterance to the more difficult task of creating new sentences by concatenating pieces of speech. The main difficulty has to do with changing the fundamental frequency; it turns out that the predictor equations, in the autocorrelation form, do not estimate formant frequencies and bandwidths accurately. This is no problem if one uses the same f_0 during resynthesis because the error is undone, but if a new f_0 is employed, the first formant may be in error by plus or minus 8% or more (Atal and Schroeder, 1975; Klatt, 1986a), and formant bandwidths can be seriously deviant. Additional losses to naturalness occur if lengthening or shortening of a segment does not quite produce the right vowel quality, or if smoothing at segment boundaries results in too rapid a change in synthesis parameters. Finally, the advantages of multipulse excitation with respect to naturalness more or less disappear in text-to-speech applications. Considering all of these limitations, it is my opinion that linear prediction resynthesis at f_0 values other than in the original recording may not have the potential quality of a formant synthesizer controlled by rule.

Other analysis–synthesis procedures have also shown an ability to reproduce speech with considerable fidelity. It has even been possible to mimic a high-pitched female singing voice by summing together, for each period, formant-like damped sinusoid waveforms that are time-windowed in such a way as to prevent superposition effects between periods (Rodet, 1984). Again, the problem with any synthesis-by-rule effort based on this type of waveform representation will be to preserve naturalness as rules are developed to create sentences in terms of the primitives of the representation.

This section on speech synthesizer models has come to four main conclusions: (1) modern formant synthesizers of several different configurations are capable of imitating many male speakers nearly perfectly, (2) some of the simplifications in a formant synthesizer result in unsatisfactory imitations of breathy high-pitched vowels that frequently occur adjacent to voiceless consonants in the speech of women and children, (3) linear prediction analysis/resynthesis is a powerful method for duplicating an utterance with high fidelity, but there are limitations on its applicability to general text synthesis, and (4) an articulatory model is likely to be the ultimate solution to the objective of natural intelligible speech synthesis by machine, but computational costs and lack of data upon which to base rules prevent immediate application of this approach.

B. Acoustic properties of phonetic segments

In order to generate speech using, e.g., a formant synthesizer, it is necessary to develop rules to convert sequences of discrete phonetic segments to time-varying control parameters. Such rules depend on data obtained by acoustic analysis of speech. Perceptual data establishing the sufficiency or relative potency of individual acoustic cues are also of considerable value in determining a rule strategy. Therefore, we first review briefly the development of a body of knowledge concerning the acoustic–phonetic characteristics of the phonetic segments of English. Many of the references to be cited appear in the Lehiste (1967) reprint collection.

The investigation of acoustic cues having the greatest importance for different speech sounds began with the use of the sound spectrograph machine at Bell Telephone Laboratories (Koenig *et al.*, 1946; Potter, 1946; Potter *et al.*, 1947; Joos, 1948). The machine produced acoustic pictures of speech. The most useful type of picture for phonetics research was the broadband sound spectrogram—an example of which is shown in Fig. 15. A broadband spectrogram is a plot of frequency versus time in which blackness represents the energy present within a 300-Hz bandwidth, as averaged over about 2–3 ms. The display was designed to represent formants as slowly changing horizontal dark bands, and to indicate f_0 as the inverse of the temporal spacing between vertical striations (at least for low-pitched voices).

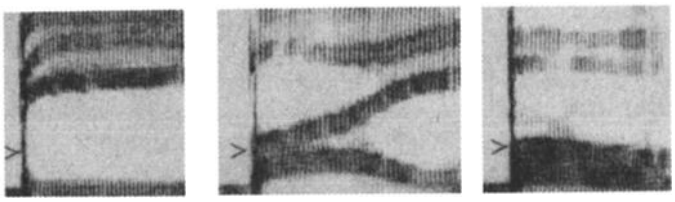

FIG. 15. Broadband spectrograms of the consonant /b/ before several vowels, illustrating the method for identifying the "hub," or the frequency from which the second formant appears to originate at consonantal onset, after Potter *et al.* (1947). The authors identified a constant hub of about 1000 Hz for /b/, although it is unclear whether this accurately reflects the onset frequency of $F2$ for /bi/.

Potter *et al.* (1947) collected sets of spectrograms depicting all of the vowels and consonants of English, and suggested ways in which to interpret the patterns they observed. They created a terminology that included terms in use today such as "stop gap" and "voice bar." In attempting to extract a common property for a stop consonant before different vowels, they defined the concept of the "hub." The "hub" is the ideal value for the second formant in each consonant. According to their observations, the second formant hub was quite useful in distinguishing between consonants having different places of articulation in English (e.g., /b/ vs /d/ vs /g/). The authors observed a fairly constant hub for/b/ before different vowels, see examples in Fig. 15,[3] and for /d/, but they said the hub for/g/ was variable across vowel context.

The investigation of the perceptual importance of various acoustic cues to a given phonetic contrast began with the use of the Pattern Playback machine at Haskins Laboratories (Cooper *et al.*, 1951). Delattre, Liberman, Cooper, and their associates created stylized versions of syllables in an effort to determine the acoustic cues sufficient for the synthesis of selected phonetic contrasts. This extensive line of research culminated in a publication suggesting explicit rules for the synthesis of English speech sounds, in which Frances Ingemann collected together a body of "synthesis-by-art" knowledge that was based on experience with the Pattern Playback (Liberman *et al.*, 1959).

The research suggested the importance of formant frequencies, formant frequency motions, spectral peaks in noise bursts, and the relative timing of onsets in different frequency regions as cues for voicing, manner, and place of articulation of consonants. The researchers emphasized the encoded nature of speech (Liberman *et al.*, 1967) in that the acoustic cues to the identity of a phoneme were spread out in time so as to overlap with cues for adjacent phonemes, and the cues were context dependent—for example the same plosive burst spectrum was heard as a different consonant depending on the vowel pattern that followed (Cooper *et al.*, 1952). There appeared to be no one invariant acoustic cue signaling the presence of a given stop consonant; rather the consonantal identity would have to be inferred from the formant transitions into an adjacent vowel. The most interesting descriptive solution to this perceptual paradox was the locus theory (Delattre *et al.*, 1955), which characterized the onset frequency of the second formant motion for a consonant–vowel transition in terms of an invisible consonant locus. The locus was determined by extrapolating backward about 50 ms from observed formant transitions for a given consonant before various vowels, Fig. 16. The importance of a virtual

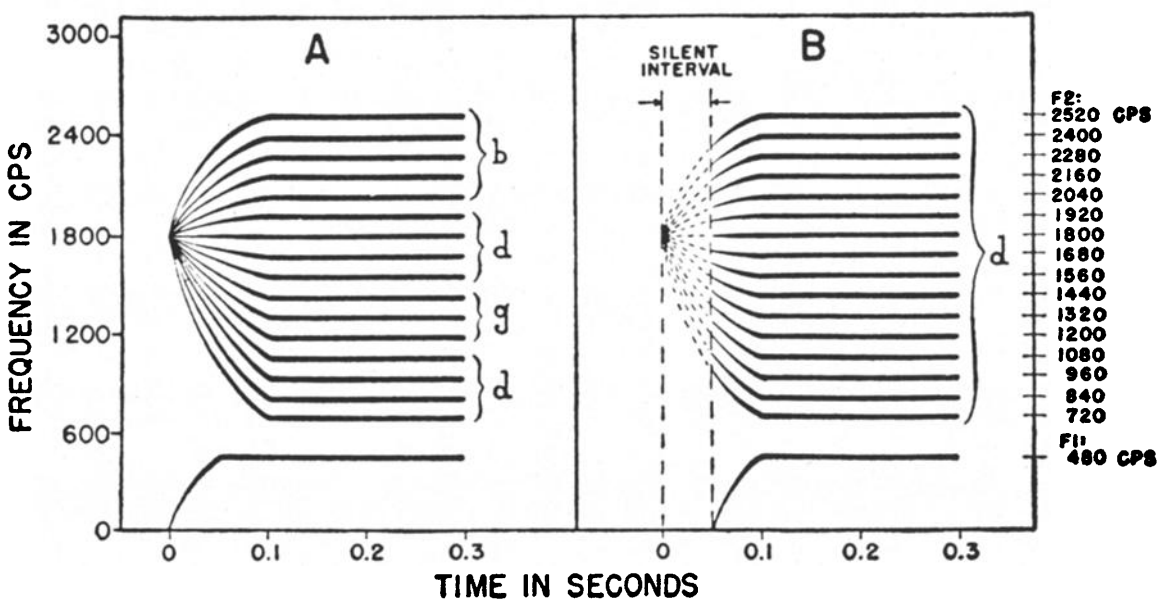

FIG. 16. The locus theory, illustrated in the right panel for the consonant /d/ before a series of vowels having the same $F1$, states that the second formant transition appears to originate from an invisible locus at 1800 Hz, after Delattre *et al.* (1955). If the second formant onset frequency (hub) is fixed at 1800 Hz, left panel, several different consonants are heard.

locus, rather than the constant $F2$ onset frequency or "hub" of Potter *et al.* (1947) was proven by synthesis of CV syllables with both types of transitions. Delattre *et al.* (1955) found that if the second formant actually started at 1800 Hz in each case, rather than at values shown in the figure, listeners heard /bi, da, gu/ instead of the intended/di, da, du/. Only when the virtual loci were employed did subjects hear/d/ in each case. The locus theory required postulation of two loci for [g], one before front vowels (where [g] is really more palatal in articulation) of 3000 Hz, and a much lower locus before back vowels. Another important observation was that phonemes sharing features such as those specifying place of articulation often shared certain acoustic patterns, making it possible to state synthesis rules efficiently in terms of familiar phonetic features, Fig. 17. Based on his experience with the Pattern Playback, Pierre Delattre became quite good at drawing stylized patterns for arbitrary sentences (example 15 of the Appendix).

a. Vowels. The acoustic theory of vowel production (Chiba and Kajiyama, 1941; Fant, 1960; Stevens and House, 1961) showed that vowels can be represented by an all-pole vocal tract transfer function, and that the relative amplitudes of the formant peaks can be predicted from a knowledge of formant frequencies, as long as the vowel is not nasalized. Peterson and Barney (1952) collected systematic data on formant frequencies and amplitudes from a wide sampling of men, women, and children. From these and many other data collection, synthesis, and perceptual validation efforts, we know that English vowels can be described in terms of the frequencies of the lowest three formants, any frequency motions associated with diphthongization (Holbrook and Fairbanks, 1962), and differences in vowel duration. Formant bandwidths also differ slightly among vowels (the best data for synthesis purposes appear to be Stevens and House, 1963); attention to details such as these is likely to lead to a slightly more natural voice quality.

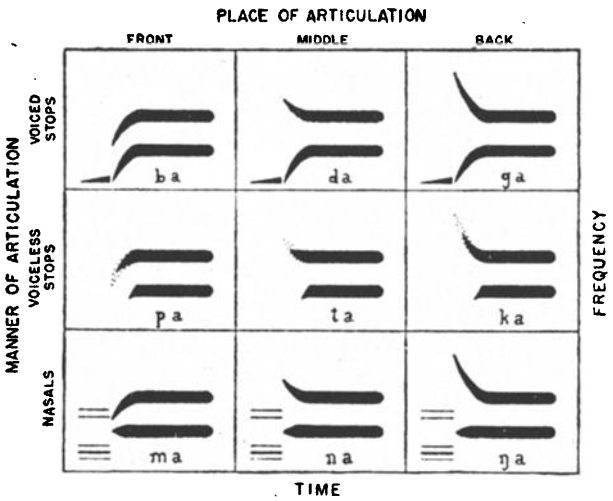

FIG. 17. Stylized pattern playback spectrograms for plosives and nasals, illustrating common $F2$ transition shapes for the same place of articulation, and common $F1$ behavior for consonants sharing the same manner of articulation, after Liberman *et al.* (1959).

b. Sonorant consonants. The non-nasal sonorant consonants of English, /w,y,r,l/, are similar to vowels, but are shorter in duration, somewhat more extreme in articulation, and are said to involve more rapid transitions into adjacent sounds than do vowels (O'Connor *et al.*, 1957; Lisker, 1957; Lehiste, 1962). Sample broadband spectrograms of these consonants in intervocalic position are shown in the bottom row of Fig. 18. Each consonant is preceded and followed by the vowel /a/, which has been truncated at the approximate midpoint of the vowel in order to fit all English consonants onto one plot. In utterance-initial position before a vowel, sonorant consonants consist of an initial brief vowel-like steady state followed by continuous formant trajectories into the following vowel. The /l/ is both sonorant and stop-like in characteristics—having a very rapid small rise in $F1$ and $F2$

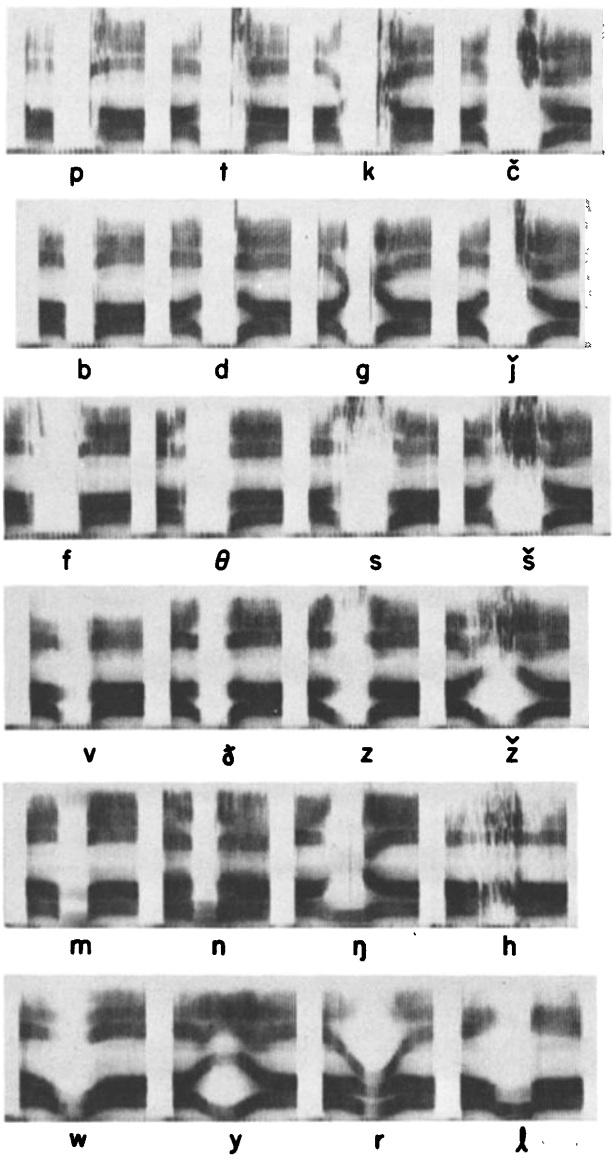

FIG. 18. Representative broadband spectrograms of English consonants produced intervocalically, i.e., in [aC'a] context. The consonant and only a portion of the vowel on each side have been excised from the full recording in order to be able to present all of the consonants in a single figure. For a brief explanation of the acoustic cues characteristic of each consonant, see text.

at the moment of release of the tongue tip from the roof of the mouth. Sonorant target values for $F1$, $F2$, and $F3$ depend somewhat on the following vowel, and a sonorant, particularly a postvocalic sonorant, can modify the formant values of the vowel a great deal (Lehiste, 1962).

The consonant /h/ is sometimes grouped with the fricatives because it is noise-excited, but /h/ functions more like a voiceless sonorant consonant. The sound source for/h/ is aspiration generated near the larynx, the vocal tract assumes the shape of the following vowel, and all formants are weakly excited by the noise.

c. Fricatives. Fricative consonants involve the generation of turbulence noise at a constriction in the vocal tract (Heinz and Stevens, 1961). The noise primarily excites the formants associated with the cavities in front of the constriction (Fant, 1960; Stevens, 1972). Acoustic properties that distinguish the English fricatives from one another include the general spectral shape of the frication noise and the motions of the formants into and out of adjacent sounds, rows 3 and 4 of Fig. 18. Each fricative noise has a relatively fixed characteristic spectral shape, although there are differences observed across speakers and across phonetic environments—e.g., anticipatory lip rounding for a rounded vowel may lower the frequencies of the most prominent spectral peaks slightly. Formant motion cues, which are particularly important for distinguishing between /f/ and /θ/ (Harris, 1958), depend to a much greater extent on the vocal tract shape of adjacent vowels. The voiced fricatives of English /v,ð,z,ž/ are shorter than voiceless /f,θ,s,š/ and usually contain simultaneous voicing at low frequencies.

d. Plosives. The voiced plosives of English, /b,d,g/ consist of a closure interval, a brief burst of turbulence noise at release, and formant transitions into and out of adjacent segments (Fischer-Jorgensen, 1954; Halle *et al.*, 1957). The spectrum of the noise burst, its duration, and the motions of the formants into a following vowel have all been shown to be important perceptual cues under some circumstances (Cooper *et al.*, 1952; Delattre *et al.*, 1955). While nominally voiced,/b,d,g/ include evidence of voicing during closure, i.e., the periodic low-frequency energy known as a voicebar, only in certain phonetic environments. Devoiced allophones, as well as several other allophones that occur in specific phonetic/stress environments, are discussed in Sec. I D 4 on phonological recoding.

The voiceless plosives of English,/p,t,k/, are similar to/b,d,g/ except that there is an interval of/h/-like aspiration noise following the burst because vocal fold adduction necessary for voicing onset is delayed (Liberman *et al.*, 1958; Lisker and Abramson, 1967). Most of the formant transitions take place while aspiration is the sound source. The burst is slightly longer and more intense, and formant transitions are somewhat less distinct in voiceless plosives, making the burst a more potent cue to place of articulation.

The English affricates /č/ and /ǰ/ are usually analyzed phonetically as consisting of a plosive followed by a fricative, i.e., /tš/ and/dž/. Their observed acoustic properties, Fig. 18, generally agree with such an assumption, although the duration of frication noise is less than in a full fricative (Gerstman, 1957).

e. Nasals. The nasal consonants /m,n,ŋ/ consist of a murmur during the interval when the oral cavity is closed, and rapid transitions into and out of adjacent segments, row 5 of Fig. 18. The murmur has a complex spectrum with a strong first formant prominence at about 300 Hz. There are both poles and zeros in the transfer function, with frequency locations dependent on the length of the side-branch resonator formed by the occluded oral cavity (Fant, 1960; Fujimura, 1962). Formant transitions into adjacent segments are similar to those for the corresponding voiced plosive (Liberman *et al.*, 1954), although there is usually some degree of nasalization of adjacent segments to complicate the picture (Fujimura, 1960). The primary acoustic cue to nasalization of a vowel is the splitting of $F1$ into a pole–zero–pole complex (Stevens *et al.*, 1987). It is difficult to distinguish one nasal consonant from another if presented only with the murmur spectrum (Malecot, 1956); formant transitions appear to be somewhat more potent cues to place of articulation, although it is perhaps the relation of the onset spectrum at release to the murmur that is perceptually most important to place-of-articulation judgments (Repp, 1986).

While this brief sketch of the acoustic properties of consonant–vowel syllables has identified some of the relevant early literature, it is important to realize that the studies referenced are not always sufficiently detailed for synthesis purposes, and isolated CV syllables are far from an exhaustive inventory of phenomena that must be treated in a rule program (see later sections on allophonics and prosody). Also, prevocalic and postvocalic consonant clusters introduce additional complications. A serious worker entering this field will probably have to develop an extensive personal data base of speech materials for analysis, rule development, and perceptual validation of chosen synthesis strategies.

C. Segmental synthesis-by-rule programs

The speech copying techniques described earlier succeed, in part, because they reproduce essentially all of the potential cues present in the waveform or spectrum, even though we may not know which cues are most important to the human listener. A synthesis-by-rule program, on the other hand, constitutes a set of rules for generating what are often highly stylized and simplified approximations to natural speech. As such, the rules are an embodiment of a theory as to exactly which cues are important for each phonetic contrast.

Early rule programs have been described and compared in a good review paper prepared by Mattingly (1974), so only the highlights will be mentioned here. Techniques have been divided into three broad categories: (1) heuristic acoustic-domain rules to control a formant synthesizer, (2) articulatory rules to control a model of the larynx and vocal tract, and (3) strategies for concatenating pieces of encoded natural speech.

1. Formant-based rule programs

The first synthesis-by-rule program capable of synthesizing speech from a phonemic representation was written by Kelly and Gerstman (1961, 1964). They used a cascaded

three-formant synthesizer that was excited by either an impulse train or a noise source, and so were somewhat limited in their ability to control formant amplitudes or to approximate voiced fricatives. Nevertheless, surprisingly good speech quality was produced by rule (example 16 of the Appendix) (with the caveat that durations and fundamental frequency contour were copied from natural speech, some hand-editing of rule output was permitted, and a familiar passage was spoken). Details of the program were never published, but rules appear to have been based on Gerstman's considerable experience with the Haskins Laboratories group (Mattingly, 1968, pp. 40–42).

Shortly thereafter, another system, both elegant in its simplicity and remarkable in its performance was created by Holmes *et al.* (1964). This publication contains a description of a parallel formant synthesizer and a complete listing of the rules and tables for synthesizing British English phonemes. The authors used a fairly simple parameter generation algorithm, whose operation was determined entirely by values in tables. A ranking procedure implemented a version of the locus theory, and allowed consonantal formant transitions to impinge on vowel target frequencies in such a way that formant undershoot of the target occurred for short vowels, as illustrated in Fig. 19. The speech quality and intelligibility of this pioneering program is remarkably good—probably better than many of the inexpensive products now on the market (example 17 of the Appendix). Unfortunately, intelligibility data for the system were never collected.

An adaptation to American English, including rules for prediction of segment durations and fundamental frequency contours, was described by Mattingly (1966, 1968) (example 20 of the Appendix). Mattingly used formant transition curves that were "S-shaped" and thus more like natural data than are linear transitions, but he found there to be little if any perceptual difference between the two types of interpolation. Allophone rules were also added at this time to permit context-conditioned modifications to table values as needed.

The Mattingly rules were combined with a set of letter-to-sound rules and a 140 000-word Kenyon and Knott phonemic dictionary, obtained from June Shoup of the Speech Communication Research Laboratory, to create an experimental Haskins text-to-speech system (Cooper *et al.*, 1973; Nye *et al.*, 1973). The system, intended to be part of a reading machine for the blind, was tested for intelligibility and optimal speaking rate (example 26 of the Appendix). The data will be discussed and compared with data for other systems in Sec. IV. Unfortunately, this pioneering effort was not pursued due to a funding lapse (Cooper *et al.*, 1984), and the device was never produced in quantity for the intended users.

Synthesis-by-rule programs proliferated during the late 1960s and early 1970s. Rabiner (1968) and Liljencrants (1969) investigated the advantages of using a critically damped second-order smoothing filter to constrain formant frequencies to move continuously in time, as required by acoustic theory. The smoothing time constant was varied depending on segmental characteristics in order to approximate the various rates of formant motion observed in natural speech. Rabiner's rules were able to synthesize CV and VC

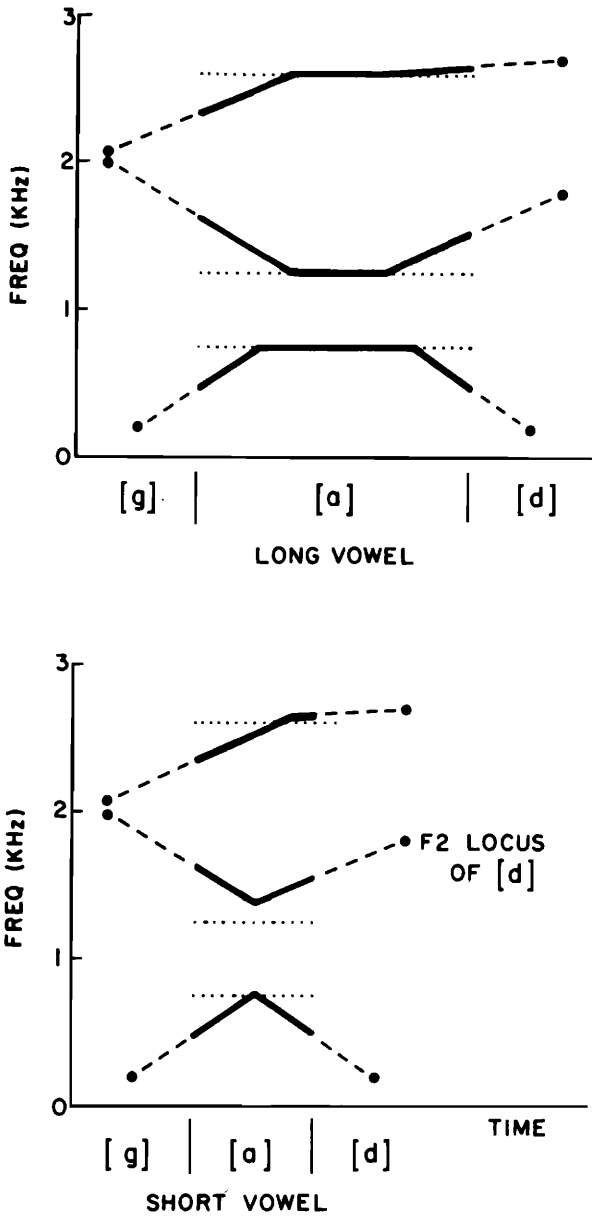

FIG. 19. According to Holmes *et al.* (1964), specification of formant motions for a simple vowel, such as the [ɑ] of "God," is a two step process: (1) Select target values for the vowel, dotted lines, and (2) use the locus theory, dashed lines, to compute smooth straight-line transitions from adjacent consonant loci toward the vowel target, solid lines. If the vowel is short, the target formant positions may never be reached (lower panel).

nonsense syllables with consonantal intelligibility of about 75%. However, when listening to recordings of a human subject producing consonant–vowel nonsense syllables, listeners are capable of much higher recognition performance, better than 99% correct (Pisoni and Hunnicutt, 1980). This represents an upper bound or goal for all rule programs attempting to synthesize speech.

Klatt (1970) extended this earlier work by formulating rules for generating CVC syllables with greater fidelity to measured characteristics of English consonants. Using a hybrid cascade/parallel formant synthesizer (Klatt, 1980) and a rule program that allowed specification of targets and

straight-line transitions á la Fig. 19, he achieved a consonantal intelligibility of 95% for CVC nonsense syllables played to trained phoneticians. Klatt had greatest difficulty with stop consonants. He, along with many others (Fant, 1973; Kewley-Port, 1982) found that the locus theory was an oversimplification that applied, at best, to two-formant acoustic patterns. Based on extensive data from a single speaker (examples are shown in Fig. 20) he tried to determine whether a modified locus concept could be created, or whether a list was needed to tabulate the starting frequencies for $F1$, $F2$, and $F3$ before each vowel. A locus theory is manifested in Fig. 20 when all of the data points lie on a straight line, i.e., when one can predict the onset frequency $F2_{\text{onset}}$ from the vowel target frequency $F2_{\text{vowel}}$ by an equation of the form:

$$F2_{\text{onset}} = F2_{\text{locus}} + k*[F2_{\text{vowel}} - F2_{\text{locus}}], \qquad (1)$$

where the locus frequency F_{locus} and degree of vowel coarticulation at the instant of release k are parameters to be fit to the observed data from each consonant.[4] At first it seemed there was little hope for resurrecting the locus concept because, as noted by Fant (1973), many complex factors cause the locus idea to fail. A transition can have both a rapid and a slow component, due to rapid release of the obstruction followed by gradual tongue body movements; a preceding vowel can influence the observed $F2$ onset of a CV transition (Öhman, 1966); and $F2$ can be relatively insensitive to oral constrictions when it is essentially a back cavity resonance, as in the vowel [i]. Klatt hypothesized that the primary influences of the vowel on consonantal articulation were fronting/backing of the tongue body and lip rounding. He therefore divided the set of English vowels into {+FRONT}, {+ROUND}, and the remainder which were {−FRONT, −ROUND}, and found that within each set, the data were sufficiently regular to be approximated by straight lines, as in Fig. 20 (Klatt, 1979b). While some data points lie slightly off the straight lines and might be better synthesized by a table look-up strategy, the recognition score of 95% correct obtained for synthetic plosives in CV nonsense syllables (Klatt, 1970) is encouraging.

Examples of burst spectra obtained from one talker, Fig. 21, support the Klatt strategy of dividing the data into vowel subsets by showing remarkably constant spectral shape and amplitude for the burst before all vowels in a given vowel set, but substantial differences across vowel sets [recall also the Cooper *et al.* (1952) perceptual results]. Burst spectra were synthesized by a strategy of selecting from one of three possible synthetic bursts depending on the following vowel. It was also noted that the properties of the burst spectra conformed to theoretical predictions concerning the quantal nature of place of articulation (Stevens, 1972), so that only formants corresponding to resonances of the cavity in front of the constriction were strongly excited by noise. For example, in [k] and [g] bursts, the noise excited $F2$ and $F4$ before back vowels, and $F3$ and $F5$ before front vowels.

One question that concerned early researchers was whether there might exist a stylized version of synthetic "super speech" that was more intelligible than natural speech because, e.g., the formant peaks were enhanced or burst spectra were "cleaned up" so as to contain only one major

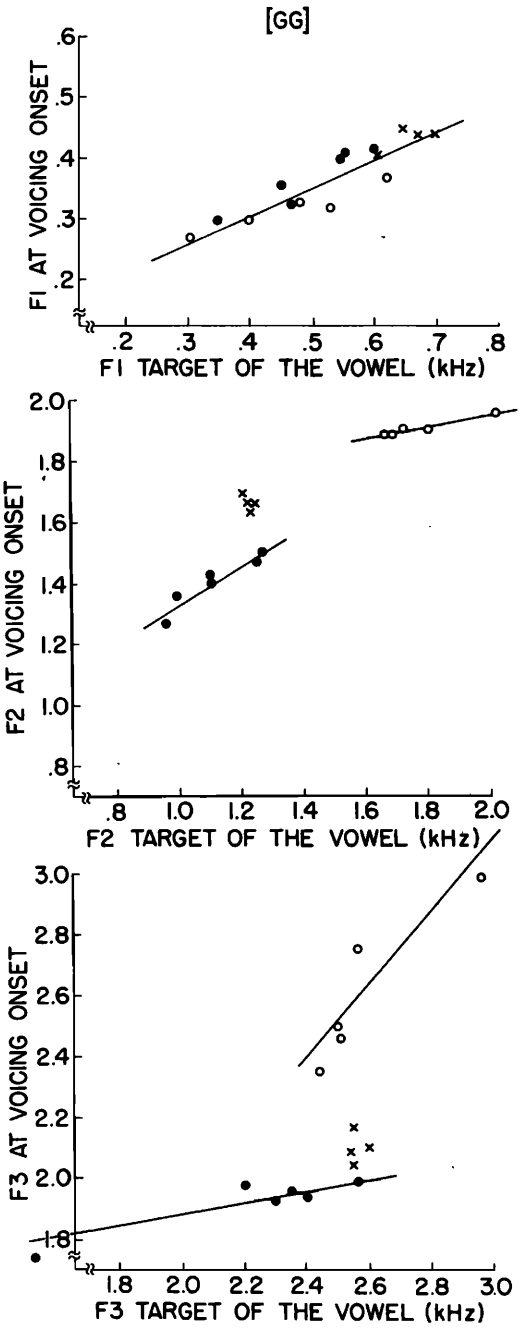

FIG. 20. The onset frequency of the formant transition into a vowel is plotted as a function of the vowel target frequency for 16 vowel nuclei before /g/, after Klatt (1979b). Each data point is an average of six tokens from a single speaker. Open circles indicate front vowels, crosses indicate back unrounded vowels, and solid circles indicate rounded vowels. Within each vowel class, data points are well fit by a straight line, confirming the existence of a locus theory equation, see text.

energy concentration, or formant transitions were more extreme than normally observed. Such efforts have always failed; synthesis that is a better match to observed natural data has always sounded better and has been measurably more intelligible. Every potential cue (acoustic regularity associated with a phonetic gesture) that has been examined has been shown to have some perceptual cue value (Liber-

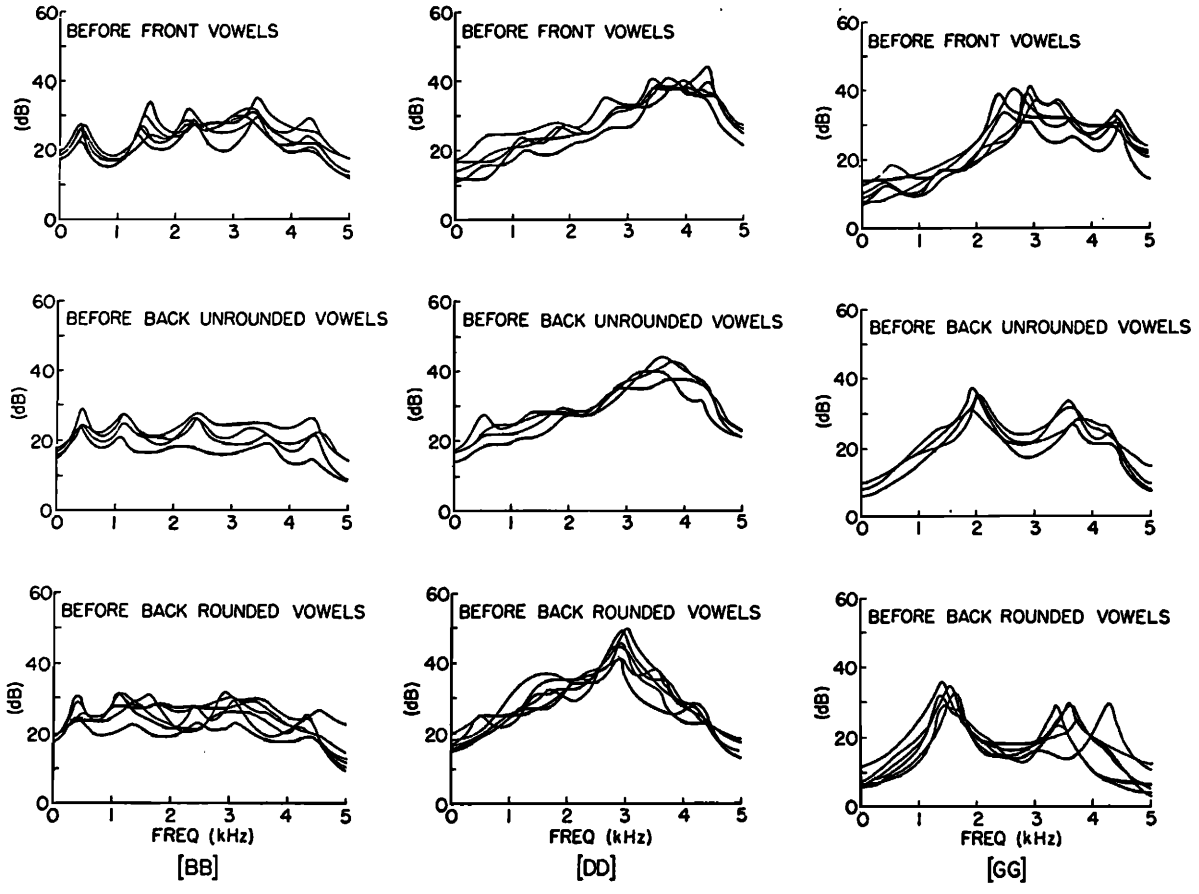

FIG. 21. Linear prediction spectra of the plosive burst for /b/, /d/, and /g/ before 16 different vowel nuclei. Each tracing is the average of six tokens of a CV syllable spoken by a single talker, after Klatt (1979b). Burst spectra for [d] and [g] display systematic changes associated with anticipatory lip rounding for a following rounded vowel, lower panels, and [g] appears to be palatalized before front vowels, top right panel.

man and Mattingly, 1985; Lisker, 1978). For example, Carlson *et al.* (1972) synthesized /g/ bursts with either a single compact energy concentration near $F2$, or $F2$ excitation *plus* a weak secondary energy concentration near $F4$ (the next front cavity resonance). They obtained best intelligibility scores using the more complicated burst that better matched natural bursts. Some cues are of course more powerful than others, but the listener appears to be responsive to an incredible number of acoustic details and performs best when the synthesis contains all known acoustic regularities (Dorman *et al.*, 1977).

The early Klatt rules for segmental synthesis were augmented by a sentence-level phonological component (Klatt, 1976b) that derived segment durations, f_0 contour, and allophonic variation by rule (example 21 of the Appendix). The program has evolved over the last 10 years, and has spawned several progeny. The 1976 version was incorporated into the MITalk text-to-speech system that was being developed in the 1970s at M.I.T. under the guidance of Jonathan Allen (Allen *et al.*, 1987). The fundamental frequency algorithm of Klatt (1976b) was replaced by one developed by O'Shaughnessy (1977). MITalk text analysis routines included a morpheme dictionary (Allen, 1976), letter-to-sound rules (Hunnicutt, 1976), and a phrase-level parser. The MITalk system evolved until 1979 when the project was terminated (Allen *et al.*, 1979; Allen *et al.*, 1987) (example 30 of the Appendix).

In 1976, the MITalk letter-to-phoneme rules (Hunnicutt, 1976) and the Klatt phoneme-to-speech program were licensed to Telesensory Systems, Inc. for incorporation into a reading machine for the blind (Goldhor and Lund, 1983). After considerable effort to transform the code into a real-time device, Telesensory Systems sold off their speech synthesis division to a newly formed company, Speech Plus, Inc. Following further development, Speech Plus came out with the Prose-2000 text-to-speech system in 1982 (Groner *et al.*, 1982) (example 32 of the Appendix). Since that time, the segmental synthesis rules have been modified to improve intelligibility over limited bandwidth long-distance telephone lines (Wright *et al.*, 1986). For example, some noise bursts and frication spectra were enhanced slightly with respect to normal levels in order to compensate for the frequency response and noise characteristics of the phone. Particular attention was paid to postvocalic consonants, where they found that adding very brief releases into a weak schwa-like element before silence ("man" = [m æ n ᵊ] improved the