Editor's note: This is the ninety-sixth in a series of review and tutorial papers on various aspects of acoustics.

Review of text-to-speech conversion for English

Dennis H. Klatt
Room 36-523, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 5 May 1986; accepted for publication 1 May 1987)

The automatic conversion of English text to synthetic speech is presently being performed, remarkably well, by a number of laboratory systems and commercial devices. Progress in this area has been made possible by advances in linguistic theory, acoustic–phonetic characterization of English sound patterns, perceptual psychology, mathematical modeling of speech production, structured programming, and computer hardware design. This review traces the early work on the development of speech synthesizers, discovery of minimal acoustic cues for phonetic contrasts, evolution of phonemic rule programs, incorporation of prosodic rules, and formulation of techniques for text analysis. Examples of rules are used liberally to illustrate the state of the art. Many of the examples are taken from Klattalk, a text-to-speech system developed by the author. A number of scientific problems are identified that prevent current systems from achieving the goal of completely human-sounding speech. While the emphasis is on rule programs that drive a formant synthesizer, alternatives such as articulatory synthesis and waveform concatenation are also reviewed. An extensive bibliography has been assembled to show both the breadth of synthesis activity and the wealth of phenomena covered by rules in the best of these programs. A recording of selected examples of the historical development of synthetic speech, enclosed as a 33 $\frac{1}{3}$-rpm record, is described in the Appendix.

PACS numbers: 43.10.Ln, 43.72.Ja

CONTENTS

INTRODUCTION

The intent of this review is to trace the history of progress toward the development of systems for converting text to speech, giving credit along the way to those responsible for the important ideas that have led to present successes. Emphasis is placed on the theory behind current algorithms. The account of this theory, in conjunction with an extensive bibliography, can serve to bring someone new to the field "up to speed" fairly rapidly, even though to some extent existing commercial systems are hidden behind a cloud of proprietary trade secrets. Perceptual data that measure the intelligibility of current systems are summarized, and a brief attempt is made to estimate the potential of the technology for practical application, especially in areas of social need. A final purpose of this undertaking is to identify the weakest links in present systems for the conversion of unrestricted

 0001-4966/87/090737-57$00.80

text to fluent, intelligible, natural sounding speech. The hope is that this critical review will focus future research in directions having the greatest payoff. The reader should be aware that the author is not an impartial outside observer, but rather an active participant in the field who has many biases that will no doubt color the review.

The steps involved in converting text to speech are illustrated in Fig. 1 (Allen, 1976). First, a set of modules must analyze the text to determine the underlying structure of the sentence, and the phonemic composition of each word. Then, a second set of modules transforms this abstract linguistic representation into a speech waveform. These processes have interesting connections to linguistic theory, models of speech production, and the acoustic–phonetic characterization of language (experimental phonetics), as well as to a topic that Vanderslice (1968) calls "synthetic elocution," or the art of effective reading out loud. The review will focus on the conversion of *English* text to speech. Systems for other languages will not be reviewed unless they have contributed to the evolution of systems for English.

It might seem more practical to store natural waveforms corresponding to each word of English, and to simply concatenate them to produce sentences, particularly considering the low cost and large capacity of new laser disk technology. However, such an approach is doomed to failure because a spoken sentence is very different from a sequence of words uttered in isolation. In a sentence, words are as short as half their duration when spoken in isolation—making concatenated speech seem painfully slow. The sentence stress pattern, rhythm, and intonation, which depend on syntactic and semantic factors, are disruptively unnatural when words are simply strung together in a concatenation scheme. Finally, words blend together at an articulatory level in ways that are important to their perceived naturalness and intelligibility. The only satisfactory way to simulate these effects is to go through an intermediate syntactic, phonological, and phonetic transformation.[1]

A second problem with approaches that attempt to store representations for whole words is that the number of words that can be encountered in free text is extremely large, due in part to the existence of an unbounded set of proper names [e.g., the Social Security Administration (1985) estimates that there are over 1.7 million different surnames in their files], as well as the existence of general rules that permit the formation of larger words by the addition of prefixes and suffixes to root words, or by compounding. Also, new words are being coined every day. It was hoped that a system employing prerecorded words might spell out such items for the listener, but this has proven to be less than satisfactory. Modern systems to be described below have fairly powerful fall-back procedures to be used when an unfamiliar word is encountered.

For expository reasons, the review is organized backwards with respect to Fig. 1. Only after we have some idea of the nature of the input information required by the synthesis routines will the second section take up the analysis of text.

A. Linguistic framework

A recent trend in linguistics has been to describe a language such as English in generative terms, the goal being to specify rules for the generation of any legitimate sentence of the language (Chomsky and Halle, 1968). I have summarized and simplified this view somewhat in Fig. 2 to indicate how it might be applied to the problem of synthesis. Linguists believe that a sentence can be represented by a sequence of discrete elements, called *phonemes*, that are drawn from a small set of about 40 such sound building blocks for English (see Table IV). These abstract phonemic symbols might be thought to represent articulatory target configurations or gestures. Thus a word like "beam" consists of three phonemes, the /b/ characterized by lip closure, the vowel /i/ characterized by a high fronted tongue position, and the nasal /m/ characterized by both lip closure and opening of the velar port to the nasal passages. The psychological reality of the phoneme as a unit for representing how words are to be spoken is attested to by collections of speech errors in which phonemic exchanges are common (Fromkin, 1971). Linguists have also found it useful to be able to refer to the components or *features* of a phoneme, such as the fact that /b/ and /m/ are both + LABIAL, while only /m/ is + NASAL. Rules describing how words change pronunciation in certain sentence contexts are often stated most efficiently in terms of features.

Phoneme strings form larger units such as syllables, words, phrases, and clauses. These structures should be indicated in the underlying representation for an utterance, because aspects of how a sentence is pronounced depend on the

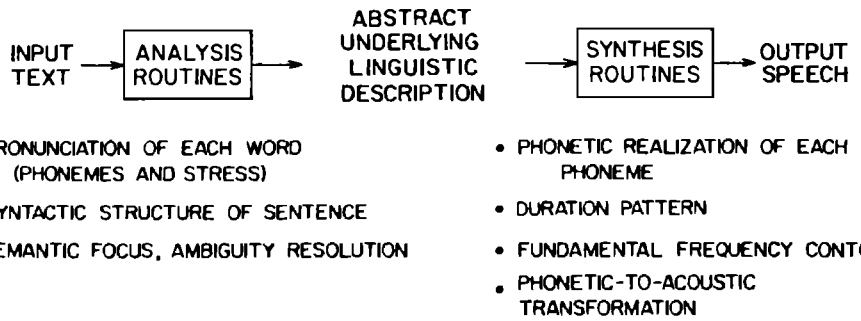

FIG. 1. Text must be converted to an abstract linguistic representation so as to be able to generate an accurate synthetic approximation to an English sentence.

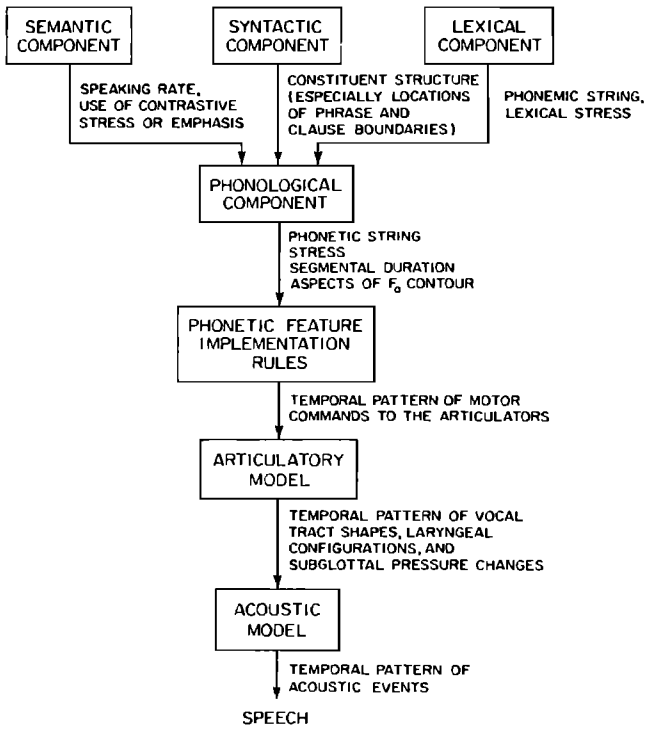

FIG. 2. Simplified block diagram of how a linguist might view the sentence generation process. An abstract linguistic representation for a sentence that is provided by the semantic component, syntactic component and lexical component undergoes various intermediate transformations before becoming an acoustic waveform.

locations of these types of boundaries. Each syllable of a word in a sentence can be assigned a strength or *stress* level. Differences in assigned stress make some syllables stand out from the others. The stress pattern has an effect on the durations of sounds and on the pitch changes over an utterance (fundamental frequency of vocal cord vibrations, or f_0). The phonological component of the grammar converts phonemic representations and information about stress and boundary types into (1) a string of phonetic segments plus (2) a superimposed pattern of timing, intensity, and f_0 motions—the latter three aspects being known as sentence *prosody*.

In mapping phonemes into sound, traditional linguists recognize a second intermediate level of representation that has been termed the phonetic segment or *allophone*. For an extreme example, the phoneme /t/ may be replaced by one of six distinctly different allophones, which will be described later in Fig. 27. The *phonological component* of the grammar includes rules to make these substitutions, either by replacing one symbol by another, or by changing the feature representation of a phoneme. The theoretical status of a phonetic level of representation (can it adequately describe individual languages and speaker behavior while simultaneously being capable of representing details in all human languages) is in some dispute, but since the text-to-speech algorithms follow allophonic substitution by other rules to make graded changes to segments, these theoretical questions are of less concern.

Unfortunately, most generative linguists have concentrated their efforts on developing rules and representational systems for the upper components of Fig. 2, and have left much of the detail concerned with articulation (feature implementation) and conversion to sound unspecified. Nevertheless, text-to-speech systems have benefited from attempts to follow this schema, and incorporate as many published phonetic details as possible within their algorithms, as we will see.

I. PHONEMES-TO-SPEECH CONVERSION

As suggested by Fig. 2, many steps are required in order to convert a phoneme string—supplemented by lexical stress, syntactic, and semantic information—into an acoustic waveform. An overview of these transformations is most easily provided by describing examples taken directly from the Klattalk algorithms (Klatt, 1982a). For example, the phonemes, stress, and syntactic symbols shown at the top in Fig. 3 for the utterance "*Joe ate his soup*" are first converted into allophones. Following the usual convention, Fig. 3 representations surround phonemes by slashes, and place square brackets around a phonetic string. Phonological rules modify three of the phonemes in this example. The /h/ of unstressed "*his*," being unstressed, is deleted, which then causes the /t/ of "*ate*" to become a flap. Finally, the postvo-

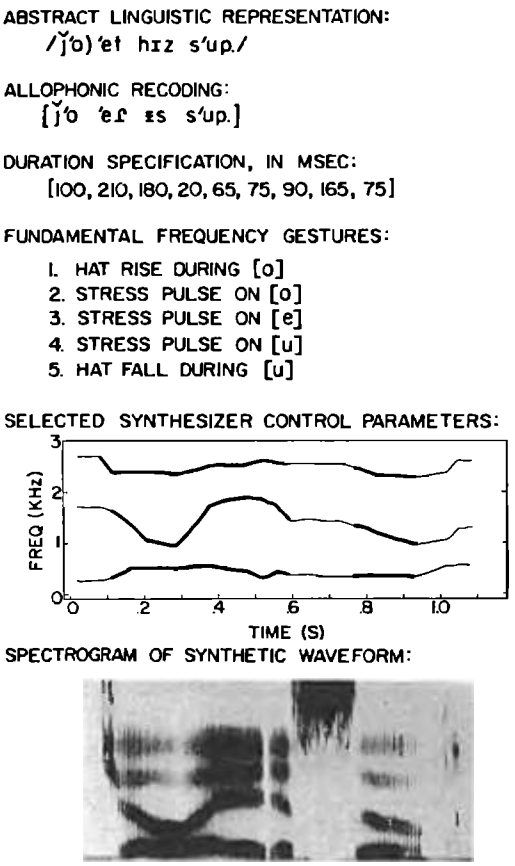

FIG. 3. An example of successive stages in the Klattalk transformation of a phonemic representation for the sentence "Joe ate his soup" to an acoustic waveform, see text.

calic/z/ of "*his*" becomes voiceless under the influence of the following voiceless /s/.[2] All other phonemes are realized in their canonical phonetic form. Of course, these canonical allophones might be modified by later rules involving stress, duration, and phonetic context, but the modifications are graded in nature and so do not call for separate discrete allophonic symbols.

Next, each phonetic segment is assigned an inherent duration by table lookup, and a set of duration rules is applied to predict changes to the duration of the segment as a function of sentential context. There are many such rules, so only a few will be illustrated. The final vowel of the sentence is lengthened by a clause-final lengthening rule. Stressed vowels are lengthened, as are the consonants that precede them in the same syllable. The vowels in "*ate*" and "*soup*" are shortened because the next consonants are voiceless. A special incompressibility constraint ensures that interacting rules cannot shorten a segment beyond a certain minimum.

Next, a fundamental frequency (f_0) contour is determined by rules that specify the locations and amplitudes of step and impulse commands that will be applied to a low-pass filter in order to generate a smooth f_0 contour as a function of time. The first rule erases the verb-phrase boundary symbol ")" in the phonemic representation because the preceding noun phrase "Joe" is too short to carry its own phrasal pattern. Then, a step rise in f_0 is placed near the start of the first stressed vowel, in accordance with a "hat theory" of intonation ('t Hart and Cohen, 1973), and a step fall is placed near the start of the final stressed vowel. These rises and falls set off syntactic units. Stress is also manifested in this rule system by causing an additional local rise on stressed vowels, using the impulse commands. The amount of rise is greatest for the first stressed vowel of a syntactic unit, and smaller thereafter. Finally, small local influences of

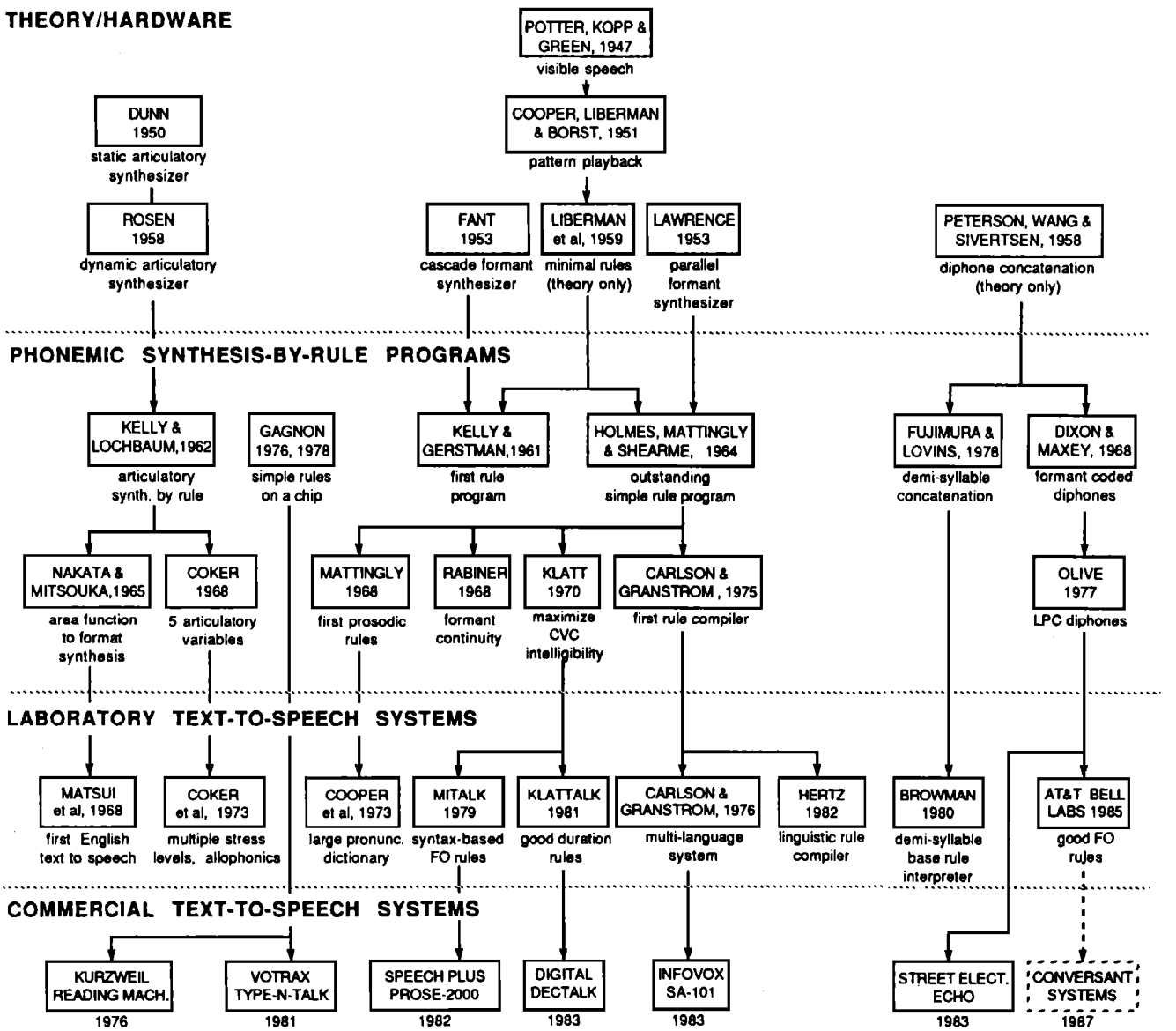

FIG. 4. Historical antecedents of the phoneme-to-speech algorithms used in several commercial text-to-speech systems.

phonetic segments are added by positioning commands to simulate f_0 rises for voiceless consonants and high vowels.

Next, a phonetic synthesis-by-rule system derives time functions that characterize the activity of voicing and noise sound sources, and the acoustic resonance properties of the vocal tract. In the Klattalk program, 19 time functions are generated, although only the three lowest formant frequency time functions are shown in Fig. 3. Rules contained in this phonetic realization module begin by selecting targets for each parameter for each phonetic segment. The target is actually a time-varying trajectory in the case of vowels because most English vowels are either diphthongs (consisting of a sequence of two articulatory targets), or include diphthongized offsets. Targets are sometimes modified by rules that take into account features of neighboring segments. Then, transitions between targets are computed according to rules that range in complexity from simple smoothing to a fairly complicated implementation of the locus theory (Delattre *et al.*, 1955; Klatt, 1979b). Most smoothing interactions involve segments adjacent to one another, but there may also be articulatory/acoustic interaction effects that span more than the adjacent segment—for example, the Klattalk program includes slow modifications to formant motions to mimic aspects of vowel-to-vowel coarticulation across a short intervening consonant (Öhman, 1966).

Finally, a formant synthesizer (Klatt, 1980) is used to convert this parametric representation into a speech waveform. The nature of the output speech waveform is illustrated by providing a broadband sound spectrogram at the bottom of Fig. 3. Klattalk might have tried to follow Fig. 2 more closely by creating a model of the articulators and a second model of the conversion of articulatory configuration to sound, but at our current state of knowledge, this was judged to be too difficult and computationally costly. Examples of attempts by others to follow an articulatory approach will be described in Sec. I C 2.

The following sections consider the various components of the synthesis-by-rule process in detail. A summary highlighting selected previous work on speech synthesis by rule is presented in block diagram form in Fig. 4. The diagram traces early work on the development of speech synthesizers, rule programs, and laboratory text-to-speech systems [many of the earlier references have been reprinted in Flanagan and Rabiner (1973)]. Several commercial text-to-speech systems are identified at the bottom of the figure (Kurzweil, 1976; Gagnon, 1978; Groner *et al.*, 1982; Bruckert *et al.*, 1983; Magnusson *et al.*, 1984), and their historical origins are suggested by the interconnecting references shown above. Other less expensive text-to-speech systems have been described elsewhere (e.g., Bell, 1983; Kaplan and Lerner, 1985).

A. Early synthesizers: Copying speech

Interest and activity in speech synthesis by mechanical and electrical devices go back a long way (Dudley and Tarnoczy, 1950); the history is well summarized by Flanagan (1972, 1976, 1981). The earliest (static) electrical formant synthesizer appears to have been built by Stewart (1922). Two resonant circuits were excited by a buzzer in this device, permitting approximations to static vowel sounds by adjusting resonance frequencies to the lowest two natural acoustic resonances of the vocal tract (formants) for each vowel.

Speech analysis/synthesis systems were conceived at the Bell Telephone Laboratories in the mid-thirties, culminating in the vocoder (Dudley, 1939), a device for analyzing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct an approximation to the original waveform. This led to the idea for a humanly controlled version of the speech synthesizer, called the "Voder" (Dudley *et al.*, 1939). The Voder, shown in Fig. 5, consisted of keys for selecting a voicing source or noise source, with a foot pedal to control fundamental frequency of voicing vibrations. The source signal was routed through ten bandpass electronic filters whose output levels were controlled by an operator's fingers. The Voder was displayed at the 1939 World's Fair in New York (example 1 of the Appendix). It took considerable skill and practice to play a sentence on the device. Intelligibility was marginal, but potential was clearly demonstrated. However, no modern text-to-speech system uses a set of fixed filter channels to create speech.

Not long thereafter, the "Pattern Playback" synthesizer was developed at the Haskins Laboratories, which permitted converting the patterns seen on broadband sound spectrograms back into sound (Cooper *et al.*, 1951; see also Young, 1948). In the Pattern Playback synthesizer shown in Fig. 6, a tone wheel generated harmonics of a 120-Hz tone, while harmonic amplitudes were controlled over time by the reflectance of painted spectrographic patterns on a moving transparent belt. Franklin Cooper, Alvin Liberman, Pierre Delattre, and their associates experimented with syllable patterns—at first copied directly from spectrograms and then simplified and stylized—in an effort to determine the acoustic cues sufficient to induce the perception of various phonetic contrasts (example 2 of the Appendix). The constant pitch made for a somewhat unnatural sound, but intelligibility was more than adequate for their purposes. In fact, words in 20 Harvard sentences were 95% intelligible if spec-

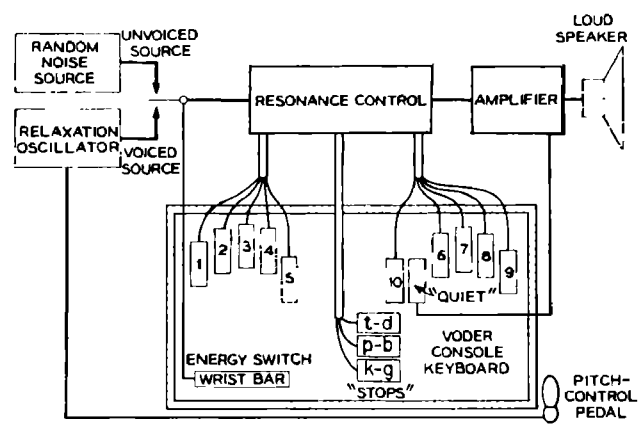

FIG. 5. The Voder speech synthesizer, consisting of a bank of filters excited by an impulse train or noise, and controlled by a piano-like keyboard, after Dudley *et al.* (1939).

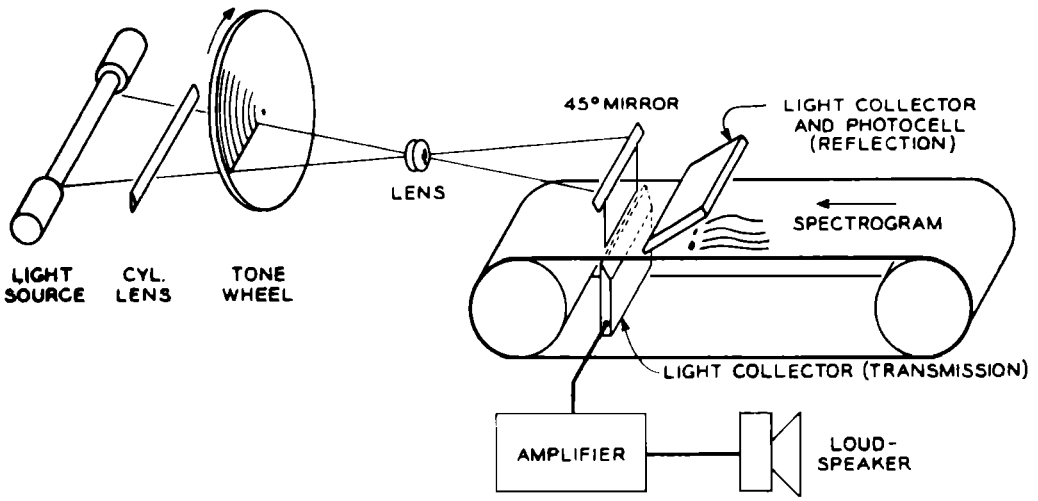

FIG. 6. The Haskins Pattern Playback, consisting of an optical system for modulating the amplitudes of a set of harmonics of 120 Hz over time depending on patterns painted on a moving transparent belt, after Cooper *et al.* (1951).

trograms were copied directly onto the transparent belt. The same words were about 85% intelligible after the spectrographic patterns had been schematized according to hypotheses about the most important aspects of observed patterns (Cooper *et al.*, 1951). Important early discoveries at Haskins are discussed in a later section.

1. The source-filter theory of speech generation

The Voder and Pattern Playback were methods for copying the time-varying spectral patterns of speech. A critical next step in the history of speech synthesis was the development of an acoustic theory of how speech is produced (summarized in Fant, 1960) and the design of formant and articulatory synthesizers based on this theory. The acoustic theory of speech production, in its simplest form, states that it is possible to view speech as the outcome of the excitation of a linear filter by one or more sound sources. The primary sources of sound are voicing, caused by the vibration of the vocal folds, and turbulence noise caused by a pressure difference across a constriction. The linear filter simulates the resonance effects of the acoustic tube formed by the pharynx, oral cavity, and lips. This vocal tract transfer function can be modeled by a set of poles—each complex conjugate pair of poles producing a local peak in the spectrum, known as a formant. At times the representation of the vocal tract transfer function in terms of a product of poles has to be augmented with zeros (antiresonators) to model the sound absorbing properties of side-branch tubes in complex articulations such as nasals, nasalized vowels, and fricatives (Fant, 1960).

2. Models of the vocal tract transfer function

Some speech synthesizers based on this acoustic theory use both poles (formant resonators) and zeros (antiformants) to model the vocal tract transfer function, while other models have tried to avoid the necessity of zeros. It has been argued that spectral notches caused by transfer function zeros are hard to detect auditorily (Malme, 1959), and therefore that the primary acoustic/perceptual effect of a zero is its influence on the amplitude of any nearby formant peak. If this assumption is true, then one may not require zero circuits in a synthesizer, as long as it is possible to adjust the amplitudes of formant peaks appropriately based on a knowledge of where the zeros of the transfer function should be. This simplification has led to a parallel formant synthesizer as one popular method for modeling the vocal tract transfer function. The outputs of a set of resonators connected in *parallel* are summed, and the input sound source amplitude of each formant resonator is determined by an independent control parameter.

The first formant synthesizers to be dynamically controlled were Walter Lawrence's Parametric Artificial Talker ("PAT") and Gunnar Fant's Orator Verbis Electris ("OVE I") (Lawrence, 1953; Fant, 1953). PAT consisted of three electronic formant resonators connected in parallel, whose inputs were either a buzz or noise. A moving glass slide was used to convert painted patterns into six time functions to control the three formant frequencies, voicing amplitude, f_0, and noise amplitude. OVE I, on the other hand, consisted of formant resonators connected in series, the lowest two of which were varied in frequency by movements in two dimensions of a mechanical arm. The amplitude and f_0 of the voicing source were determined by hand-held potentiometers. OVE I was restricted to the production of vowel-like sounds. PAT and OVE I engaged in an amusing conversation at a conference at MIT in 1956 (examples 3 and 4 of the Appendix).

Improvements were made in the synthesizers and control strategies over the next few years, so that when PAT and OVE met again on the stage at the 1962 Stockholm Speech Communication Conference, both were capable of a remarkably close approximation to a human sentence (examples 5 and 6 of the Appendix). PAT was first modified to have individual formant amplitude controls and a separate circuit for fricatives; it was later converted to cascade operation (Anthony and Lawrence, 1962). OVE I had evolved into OVE II (Fant and Martony, 1962), which included a separate static branch to simulate nasal murmurs and a special cascade of two formants and one antiformant to simulate a simplified approximation to the vocal tract transfer function

for frication noise excitation, Fig. 7. These circuits represent constraining idealizations/simplifications compared with underlying acoustic theory; it remained to be shown whether the new model was capable of synthesizing highly intelligible versions of consonants in various languages of the world.

The designers of the original PAT and OVE disagreed on whether the transfer function of the acoustic tube formed by the vocal tract should be modeled by a set of formant resonators connected in cascade (Fant, 1953, 1956, 1959, 1960) or connected in parallel (Lawrence, 1953; see also Holmes, 1973). The authors were in complete agreement as to the theory (see Flanagan, 1957, for a discussion of the mathematical relations between the two approaches) but disagreed on practical matters concerning whether it was possible to approximate vowel nasalization adequately in a cascade model, or how to avoid peculiarities in the transfer function produced by a parallel configuration when formant amplitude control settings were not perfect. The arguments persist, although at a much more sophisticated level (Holmes, 1983).

Modern synthesizers have largely abandoned electronic circuitry in favor of simulation on a digital computer (Gold and Rabiner, 1968) or construction of special-purpose digital hardware. Designs have added subtleties such as an ability to amplitude modulate the noise in a voiced fricative due to the modulation of the air stream induced by the vibrating vocal folds (Maxey, 1963; Rabiner, 1968), and have added more variable control parameters, but have otherwise not changed greatly (see references cited in Klatt, 1980). The desirability of using a *hybrid* synthesizer with cascaded formants (and an extra pole-zero pair for mimicking nasalization) for synthesis of sonorants, and parallel formants (with the same formant frequency values) for synthesis of obstruents was proposed by Klatt (1972). Klatt argued that the quantal theory of consonant place of articulation (Stevens, 1972) could be implemented directly by simple rules in such a synthesizer. The publication of this synthesizer as a Fortran listing (Klatt, 1980) promoted its use for perceptual experimentation in many laboratories, facilitating replication of stimuli and experimental results.

An important milestone in the development of speech synthesizers was the demonstration that synthetic speech could be so good that the average listener could not tell the difference between a synthetic and natural sentence when presented with both in sequence (example 8 of the Appendix). The demonstration occurred at the 1972 Boston Speech Communication Conference when John Holmes described a new version of a parallel formant synthesizer (Holmes, 1973). Holmes had spent a winter much earlier working with OVE II to synthesize a good copy of the sentence "I enjoy the simple life" spoken by a man, but had more difficulty with a female utterance (Holmes, 1961) (example 7 of the Appendix). Considering his experience with both cascade and parallel formant models, it is interesting to note that Holmes now much prefers the parallel model shown in Fig. 8 when the objective is to match a natural recording of a particular speaker. His argument, which is somewhat complex, is presented in detail in Holmes (1973, 1983). In essence, he showed that it is desirable to use a voicing waveform based on that of the speaker being modeled. This waveform can be obtained by inverse filtering vowels produced by the speaker to be imitated (the inverse filter, when properly adjusted, cancels the acoustic effects of the vocal tract transfer function). Holmes noted that stylized glottal pulses of the type used in conventional formant synthesizers work nearly as well. After adjusting the fre-

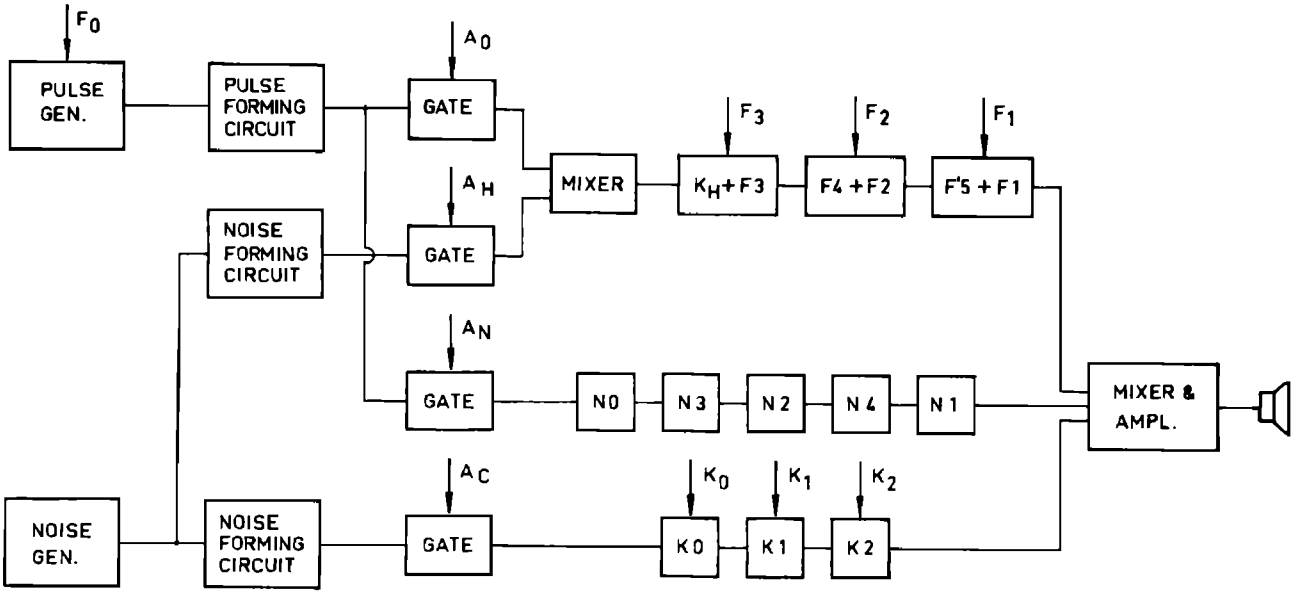

FIG. 7. The OVE II speech synthesizer, consisting of three separate circuits to model the transfer function of the vocal tract for vowels (top), nasals (middle), and obstruent consonants (bottom), after Fant and Martony (1962). Available sound sources are voicing (top), aspiration noise (middle), and frication noise (bottom).

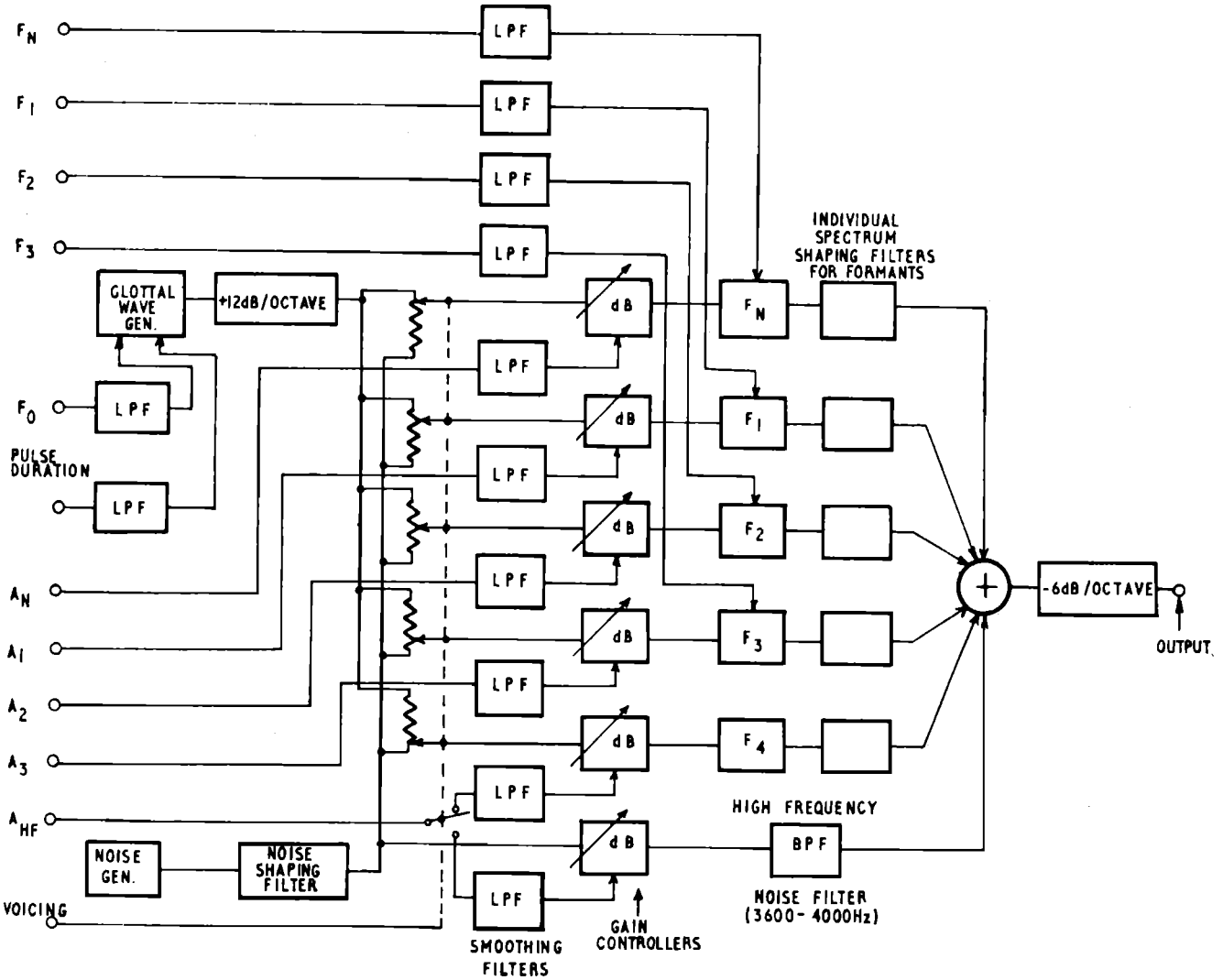

FIG. 8. The Holmes parallel formant synthesizer, consisting of four parallel formants and a nasal formant, each excited by a variable mixture of voicing and/or noise, after Holmes (1973).

quency and amplitude of the voicing source so as to mimic the fluctuations seen in the sentence, Holmes spent a long time carefully adjusting formant frequencies and amplitudes on a trial-and-error basis (see Fig. 9). He found that much of the detailed period-to-period variability in the spectra of natural speech can be mimicked by proper adjustments to the amplitudes of parallel formants—even though we may not as yet have a good enough theory and source model to account for all of this natural variation. According to Holmes, the observed irregularities in the spectrum between the formant peaks are of little perceptual importance; only the strong harmonics near a formant peak and below $F1$ must be synthesized with the correct amplitudes in order to mimic an utterance with a high degree of perceptual fidelity. Holmes also showed that phase relations among harmonics of the voicing source are important for earphone listening, but not when loudspeakers are used and the sound is modified by the reverberation of ordinary room acoustics. The Holmes synthesizer has recently been implemented on a real-time signal processing chip (Quarmby and Holmes, 1984).

Translation of the Holmes voice imitating abilities into rules for automatic synthesis of natural voice qualities has not, as yet, been successfully achieved. His parallel synthesizer is clearly up to the job, at least for male voices, so the problem remains one of developing an appropriate theory of control. Of course, it may be that the right theory will suggest a quite different model, such as an articulatory synthesizer.

3. Models of the voicing source

The voicing sound source used in a formant synthesizer has evolved from the simple sawtooth waveforms and filtered impulse train used in early designs. An impulse train filtered by a two-pole low-pass filter, displayed at the top in Fig. 10, has about the right average spectrum, but the phase of this waveform is wrong. Primary excitation of the vocal tract filters occurs at a time corresponding to the instant the folds open, rather than at closure. Furthermore, the spectrum envelope is perfectly regular (i.e., monotonically de-

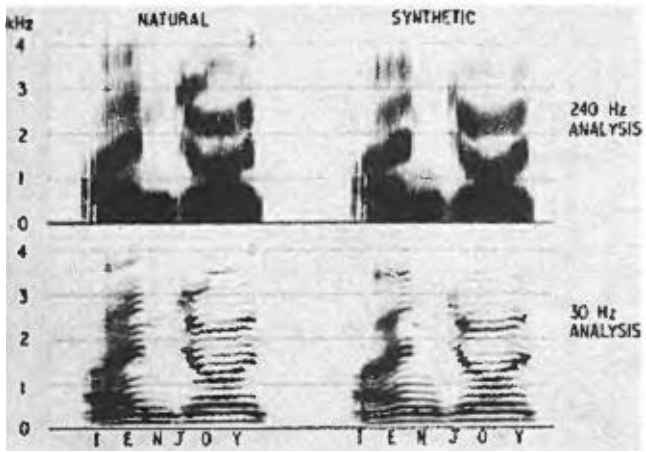

FIG. 9. Comparison of broadband and narrow-band sound spectrograms of a natural utterance and a synthetic imitation produced by Holmes (1973).

creasing at about 12 dB per octave), which contrasts with evidence indicating the presence of zeros in the spectra of normal voicing waveforms (Flanagan, 1958; Miller, 1959; Mathews *et al.*, 1961; Monsen and Engebretson, 1977; Fant, 1979; Sundberg and Gauffin, 1979; Ananthapadmanabha, 1984).

Perceptual data (Rosenberg, 1971) and theoretical considerations (Titze and Talkin, 1979) suggest ways in which the simulation of the glottal waveform might be improved. For example, Rothenberg *et al.* (1975) constructed a three-parameter model of the voicing waveform that can produce a family of more natural waveshapes varying with respect to fundamental frequency, amplitude, open quotient (ratio of open time to total period), degree of static glottal opening, and breathiness. Some of these degrees of freedom are illustrated in Fig. 10. The model is used in the Infovox SA-101 text-to-speech system (Magnusson *et al.*, 1984).

More recently, Fant *et al.* (1985) have proposed a mathematical model having similar capabilities, but with more direct control over the important acoustic variables. Some of the flexibility is illustrated in the spectral domain in Fig. 11. General spectral tilt, locations of spectral zeros, and intensity of the fundamental component are under user control. The Klattalk voicing source waveform defined in the top half of Fig. 12, which is quite similar to the Fant model, can be modified in (1) open period, (2) abruptness of the closing component of the waveform, (3) breathiness, and (4) degree of diplophonic vibration (alternate periods more similar than adjacent periods). However, rules for dynamic control of these variables are quite primitive. The limited naturalness of synthetic speech from this and all other simi-

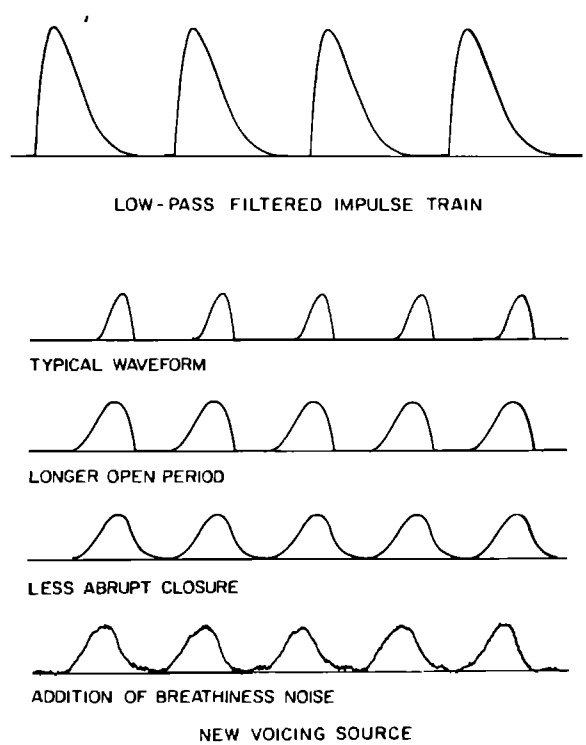

FIG. 10. Comparison of voicing waveforms consisting of a filtered impulse (top) and several more natural waveforms produced by varying the open quotient, spectral tilt, and breathiness in the Klattalk model.

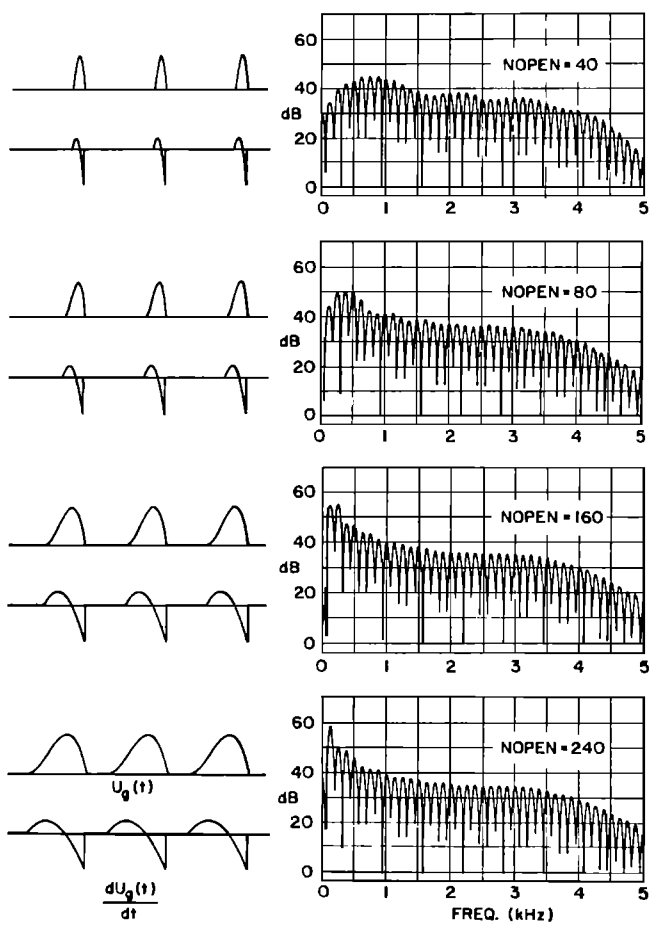

FIG. 11. Selected magnitude spectra of the output of the voicing source model of Klattalk as the duration of the open portion of the glottal cycle is varied from 1 ms (NOPEN = 40) to 6 ms (NOPEN = 240). Note the significant change in relative energy content at low frequencies; the amplitude of the first harmonic varies by about 24 dB from top to bottom panels. Short samples of the glottal waveform $U_g(t)$ and its first derivative are shown to the left of each spectrum.

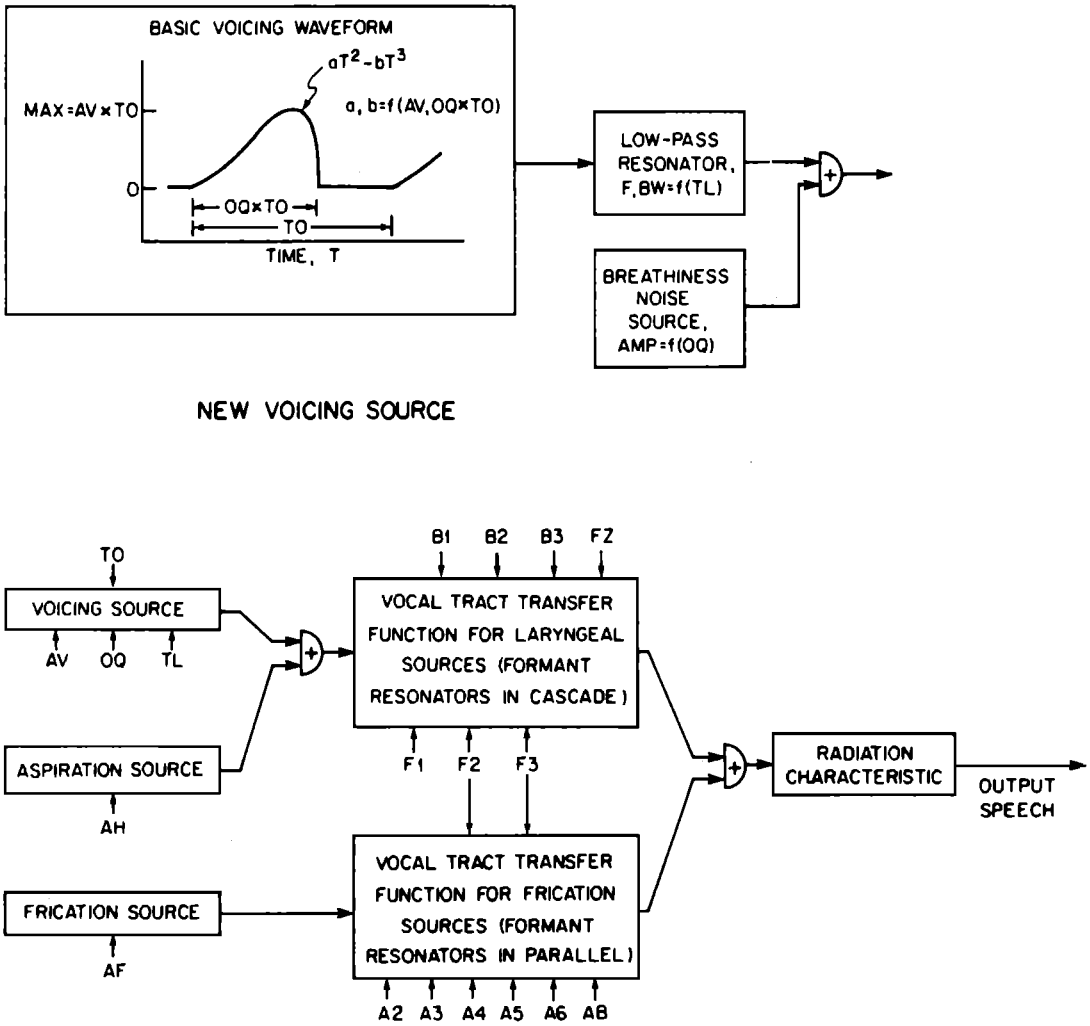

FIG. 12. Block diagram of the Klattalk synthesizer in which a new voicing algorithm (top) has been added to the synthesizer (bottom) that was described in Klatt (1980). Nineteen variable control parameters are identified, including the new voicing source parameters OQ (open quotient) and TL (spectral 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

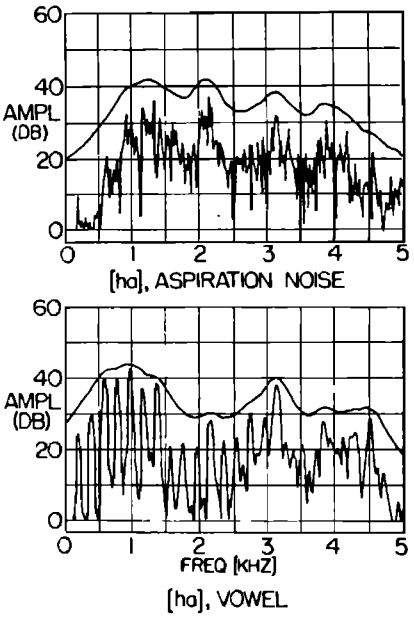

FIG. 13. The magnitude spectrum of 50 ms of aspiration noise in the syllable [ha] (top) reveals a strong subglottal formant at 2150 Hz, between $F2 = 1300$ Hz and $F3 = 3200$ Hz. Additional subglottal resonances are faintly evident at about 600 and 1600 Hz. The effect of the subglottal system on the 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

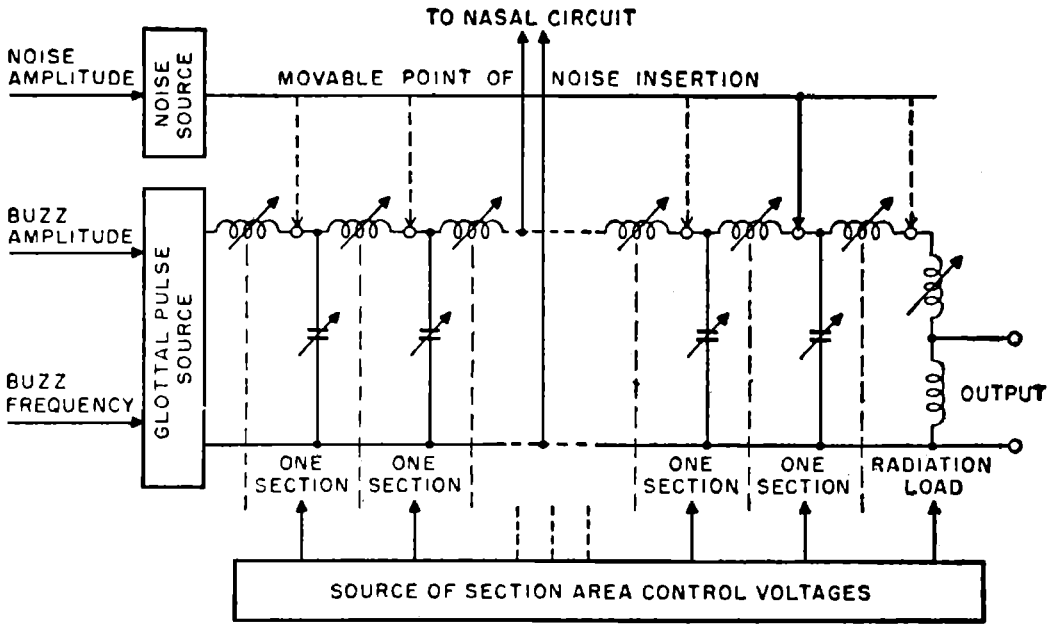

FIG. 14. The DAVO (Dynamic Analog of the VOcal tract) synthesizer, consisting of a ladder network of inductors and capacitors, each section of which mimics the properties of a short section of the vocal or nasal passages, after Rosen (1958). Inductance and capacitance values are determined by the area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

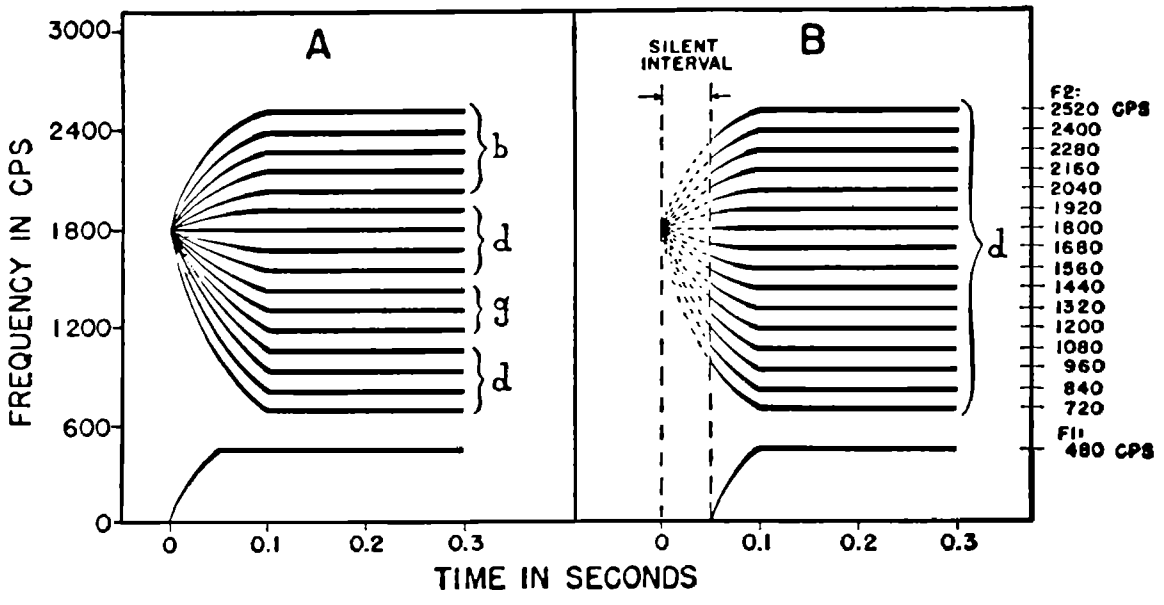

FIG. 16. The locus theory, illustrated in the right panel for the consonant /d/ before a series of vowels having the same $F1$, states that the second formant transition appears to originate from an invisible locus at 1800 Hz, after Delattre *et al.* (1955). If the second formant onset frequency (hub)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

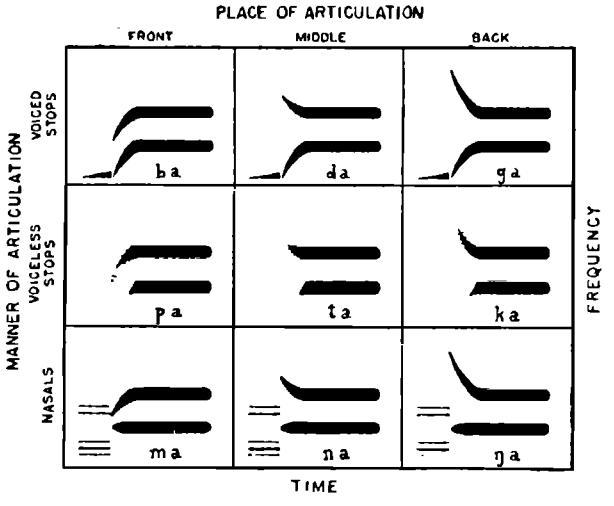

FIG. 17. Stylized pattern playback spectrograms for plosives and nasals, illustrating common $F2$ transition shapes for the same place of articulation, and common $F1$ behavior for consonants sharing the same manner of articulation, after Liberman *et al.* (1959).

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 $F2$

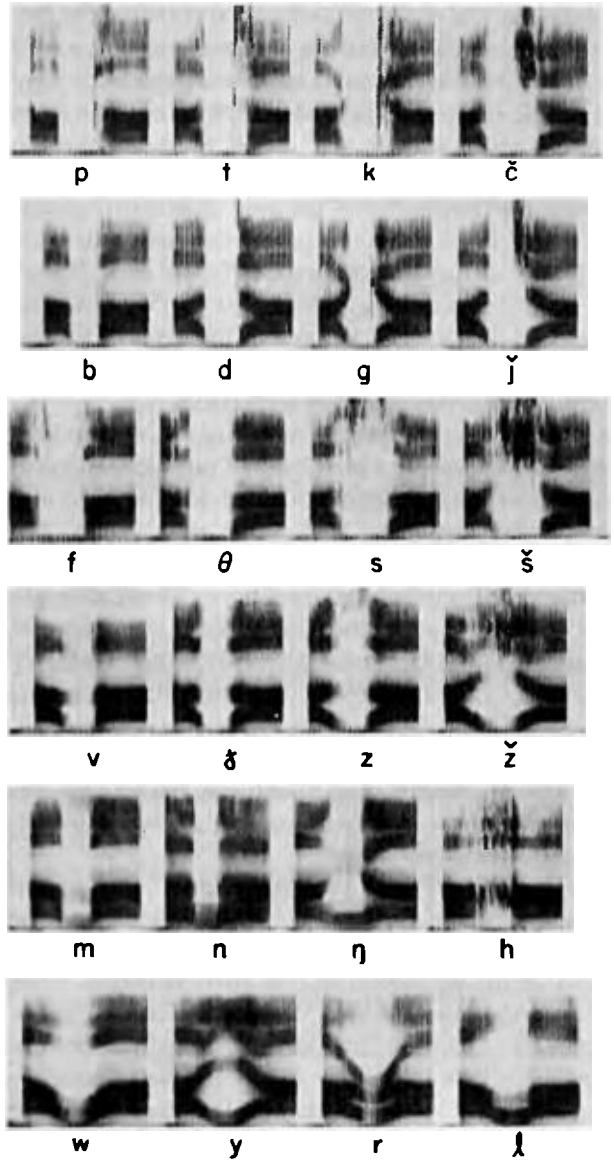

FIG. 18. Representative broadband spectrograms of English consonants produced intervocalically, i.e., in [ɑC'ɑ] context. The consonant and only a portion of the vowel on each side have been excised from the full recording in order to be able to present all of the consonants in a single figure. For a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

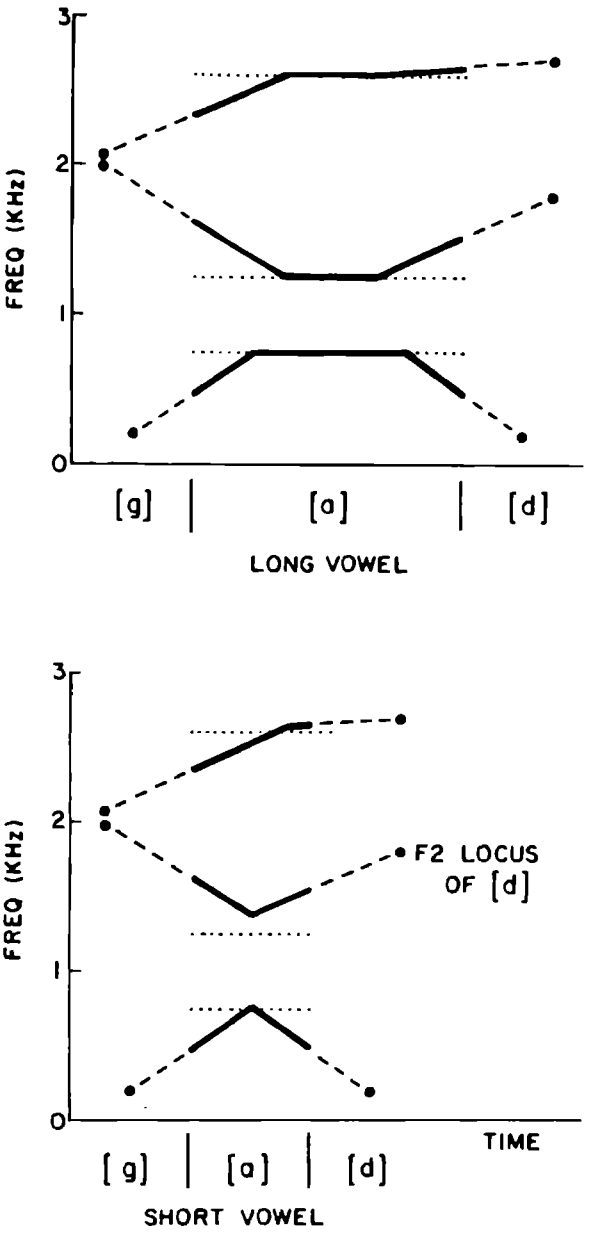

FIG. 19. According to Holmes *et al.* (1964), specification of formant motions for a simple vowel, such as the [ɑ] of "God," is a two step process: (1) Select target values for the vowel, dotted lines, and (2) use the locus theory, dashed lines, to compute smooth straight-line transitions from adjacent 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

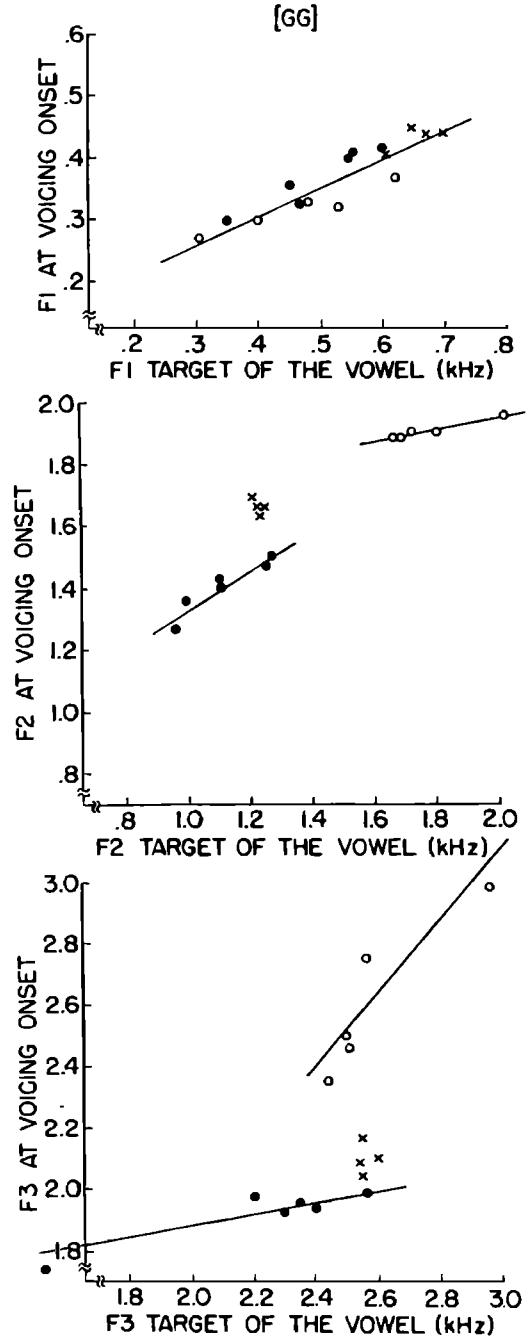

FIG. 20. The onset frequency of the formant transition into a vowel is plotted as a function of the vowel target frequency for 16 vowel nuclei before /g/, after Klatt (1979b). Each data point is an average of six tokens from a single speaker. Open circles indicate front vowels, crosses indicate back 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-

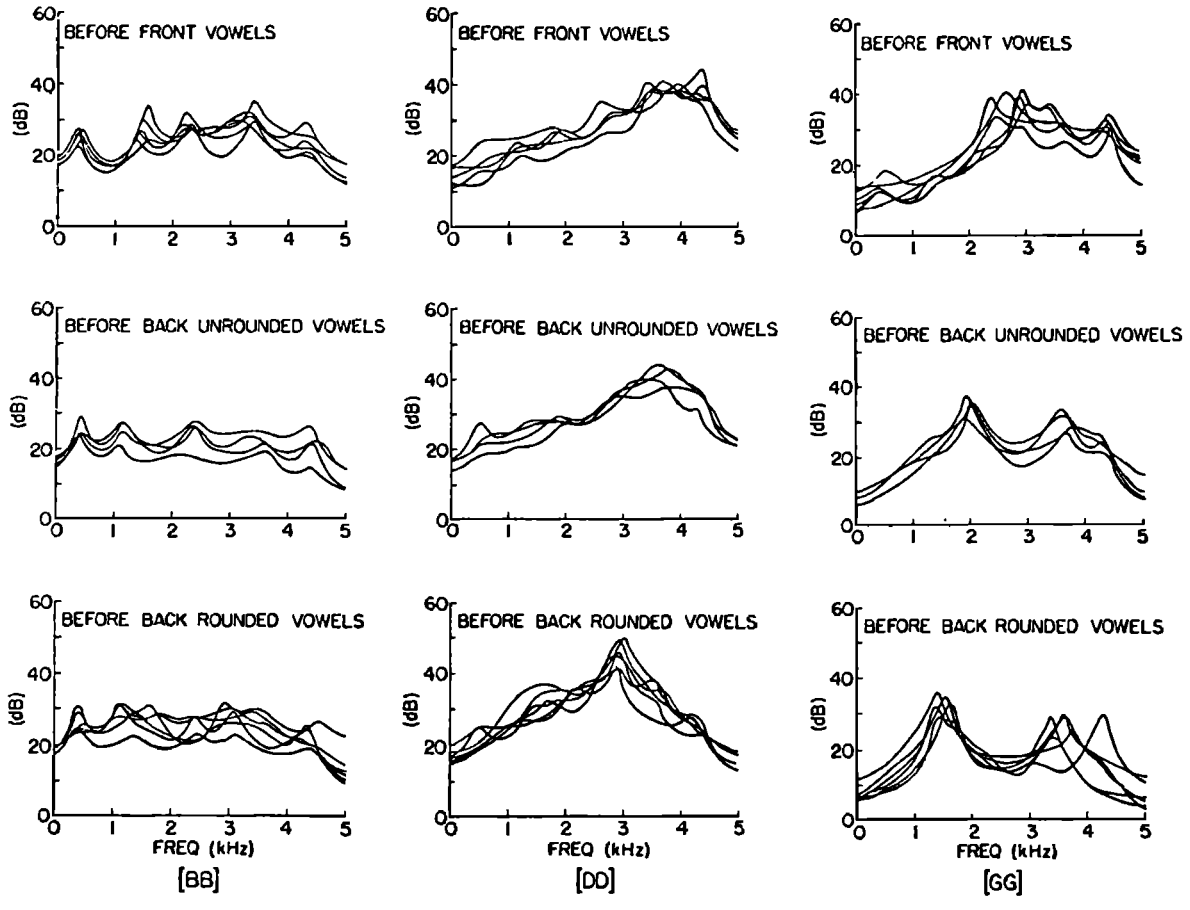

FIG. 21. Linear prediction spectra of the plosive burst for /b/, /d/, and /g/ before 16 different vowel nuclei. Each tracing is the average of six tokens of a CV syllable spoken by a single talker, after Klatt (1979b). Burst spectra for [d] and [g] display systematic changes associated with anticipatory 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

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

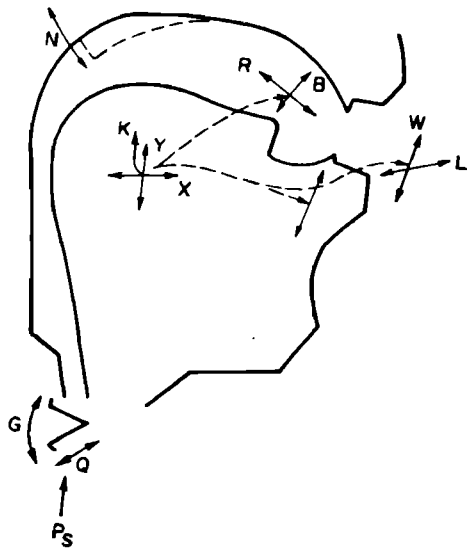

FIG. 22. A simplified low-dimensionality model of the vocal tract (Coker, 1976) permits the lips to open (W) or protrude (L); the tongue body to be raised (Y) or be backed (X) or be bunched for a velar closure (K); the tongue tip to be raised with respect to the body (B) or moved forward (R); the jaw 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).

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.

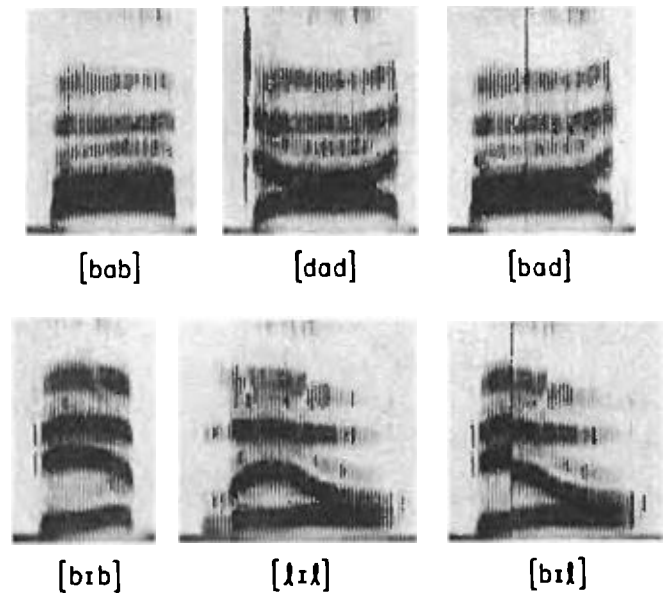

FIG. 23. Broadband spectrograms suggesting that the diphones "ba" and "ad" obtained from the syllables [bab] and [dad] can be juxtaposed to synthesize a good approximation to [bad] (upper right panel). However, synthesizing [bɪl] (lower right panel) from diphones extracted from [bɪb] and [lɪl] requires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

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" = [wɛ + ɛ^{y}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-

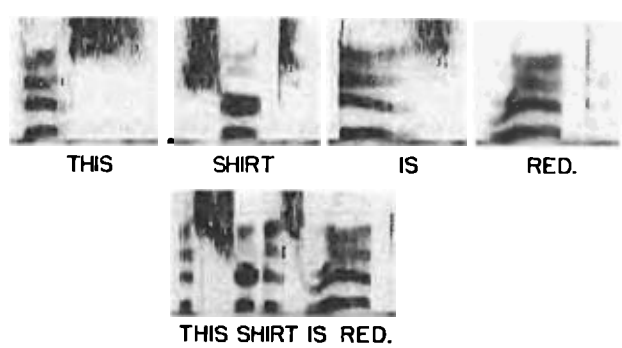

FIG. 24. Broadband spectrograms indicating that a sentence is very different from a concatenated string of words recorded in isolation. Words in sentence context are generally much shorter in duration, are subject to coarticulation at word boundaries, and undergo phonetic recoding—for example, the /t/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

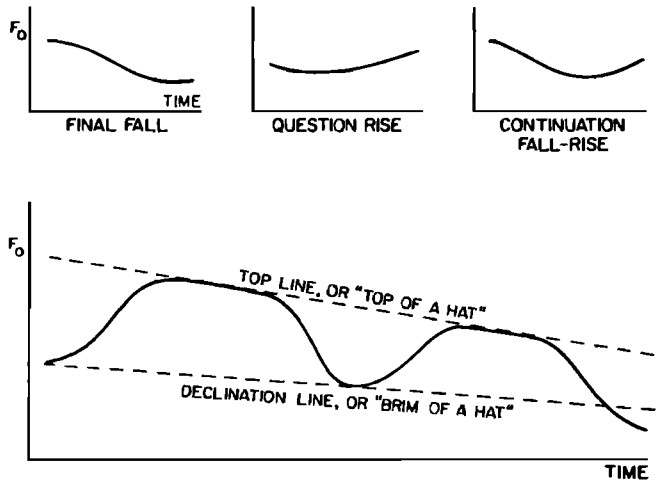

FIG. 25. Three typical clause-final intonation contours (top), and an example of a fundamental frequency "hat pattern" of rises and falls between the brim and top of a hat for a two-clause sentence (bottom).

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.

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

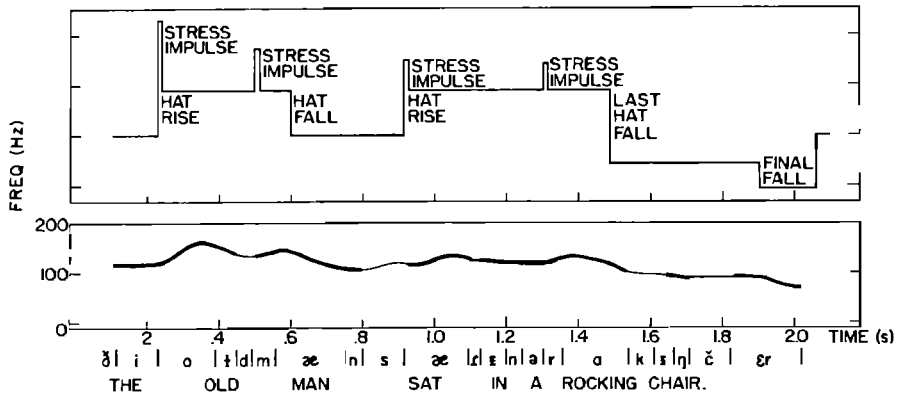

FIG. 26. The Klattalk f_0 contour shown at the bottom was generated by sending a sum of the various step and "impulsive" commands shown above through a low-pass smoothing filter. Additional small step commands associated with tongue height and glottal state, and a gradual declination line also served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

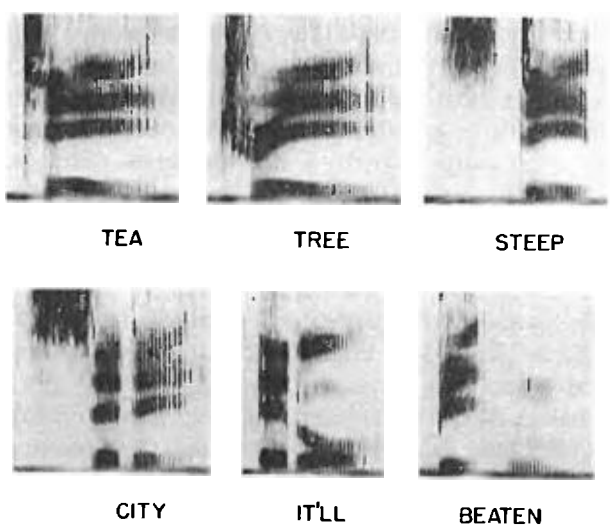

FIG. 27. Broadband spectrograms illustrating allophonic changes to /t/: a normal aspirated /t/ in "tea," affricated in "tree," unaspirated in "steep," flapped in "city," lateral or glottal release in "it'll," and nasal or glottal release in "beaten," after Zue and Laferierre (1979).

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" [dɪjʊ] 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-I^s-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 [I^s]. 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 [I^s] trajectory and the superimposed incremental trajectories for the initial and final consonants, as

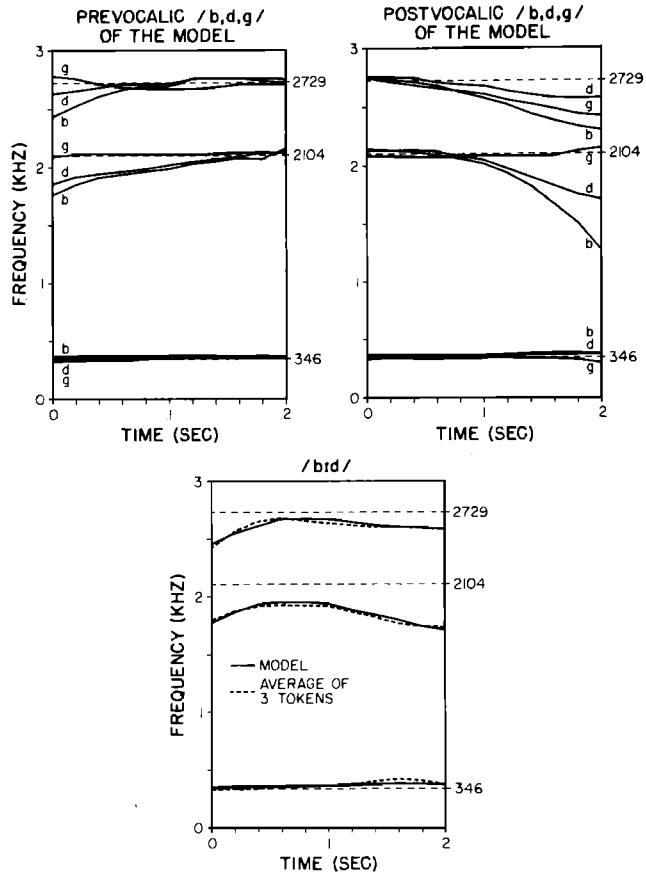

FIG. 28. Approximation to the formant transitions for the CVC syllable [bɪd] (bottom panel) was derived in terms of a model (top panels) that takes into account the average vowel formant positions (dashed lines) and incremental perturbations due to each consonant (solid lines), after Broad and Fertig (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.

shown at the bottom in Fig. 28. They were quite successful for this vowel, which is fortunate, because the general technique is essentially identical to the synthesis strategies used in Klattalk and other phonemic rule programs to generate consonant–vowel–consonant formant patterns.

These data should be augmented for other vowels, but the analysis task is formidable, so only partial data are available on some vowels in symmetrical CVC contexts (Stevens and House, 1963). Similarly, the same kinds of studies should be performed for other speakers, and at several syllable durations and degrees of stress. Of particular interest are rules that modify vowels in sentence contexts depending on consonantal context, stress, and duration. From what little is known, it appears that the vowel space shrinks when one goes from words spoken in isolation to sentences (Fant *et al.*, 1974; Shearme and Holmes, 1961), but it is not clear whether vowels tend to neutralize toward schwa, or simply accede to articulatory demands of adjacent consonants (Lindblom, 1963). It is possible that some of the subjective impression of unnaturalness and "foreign dialect" of synthesis-by-rule systems can be attributed to insufficient attention to details of this sort, both known and those yet to be discovered.

The formant transitions for a CV syllable depend to some extent on the nature of the phonetic segment that precedes the consonant. Öhman (1966) has published data on formant motions for [b,d,g] in different VCV environments that demonstrate significant interactions (Fig. 29), and Martin and Bunnell (1982) have shown that listeners expect these coarticulatory shifts—subjects show reaction time deficits when the formant shifts are not present.

Text-to-speech systems have only begun to simulate the details of the phenomena noted in this section (Coker *et al.*, 1973; Klatt, 1976b). Klattalk now contains a separate subroutine for allophonic substitution rules, as well as many detailed parameter adjustment rules in the part of the program concerned with drawing parameter values for phonetic sequences. Taken in total, these rules characterize all of the allophonic variants and word-boundary cues described here, although the rules are simplified and generalized beyond the available data in such a way that they probably do not adequately represent the environments where the rule should be

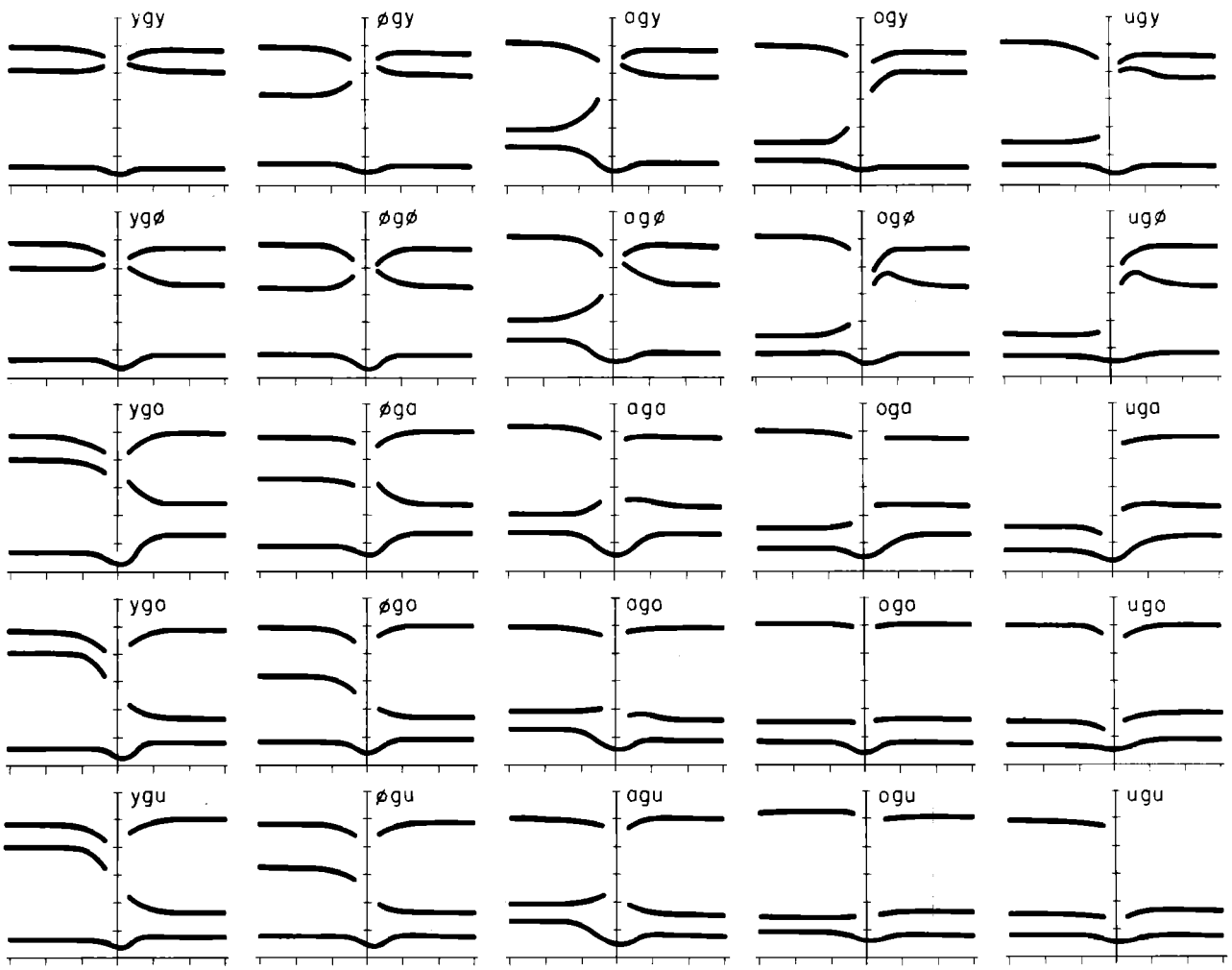

FIG. 29. Formant transitions for [g] as a function of preceding and following vowels, after Öhman (1966). Note that the formant motions for, e.g., [ga] release (middle horizontal panels) change as the preceding vowel changes.

evoked or the detailed acoustic effects of the transformation. For example, only the fronting/backing anticipation of the interacting vowels in a VCV sequence described by Öhman has been implemented as a change to the $F2$ trajectory. The list of discrete allophones inserted and manipulated by internal Klattalk rules, shown in Table III, is rather small. All other allophonic variation is created by modifying synthesizer control parameter data directly rather than by defining a discrete symbol.

In summary, it is likely that this area of allophonic detail and prosodic specification is one of the weaker aspects of rule systems, and contributes significantly to the perception of unnaturalness attributed to synthetic speech. Incremental improvements that are made to these rules on the basis of comparisons between rule output and natural speech cannot help but lead to improved performance of text-to-speech systems.

II. TEXT-TO-PHONEMES CONVERSION

Having considered the steps required to go from an abstract linguistic description to synthetic speech, we now turn to the problem of deriving this description from text. The recognition of printed characters, as required in, e.g., a reading machine for the blind, is beyond the scope of this review. We will assume that an ASCII representation of each input sentence is available as input to the text analysis module of a text-to-speech system. From considerations outlined in the previous section, it is clear that the text analysis routines have a formidable task. Ideally, the input is to be analyzed in such a way as to:

- reformat everything encountered (e.g., digits, abbreviations) into words and punctuation,
- parse the sentence to establish the surface syntactic structure,
- find the semantically determined locations of contrastive and emphatic stress,
- derive a phonemic representation for each word,
- assign a (lexical) stress pattern to each word.

For example, the input ASCII string for a typical input sentence, shown below, was processed by rules of Klattalk to derive an abstract linguistic representation consisting of phonemes, stress, and syntactic symbols. First, the word-formatting module transformed the numerals *"23"* into the words *"twenty-three."*

INPUT TEXT:
The 23 protesters were arrested.
REFORMATTED INTO WORDS:
The twenty-three protesters were arrested.
(PARTIAL) SYNTACTIC ANALYSIS:
The twenty-three protesters) were arrested.
SEMANTIC ANALYSIS:
None.
(PARTIAL) MORPHEMIC ANALYSIS:
The twenty-three protest-er-s) were arrest-ed.
PHONEMIC CONVERSION AND LEXICAL STRESS ASSIGNMENT:
/ðə tw´ɛnti θr´i pr´otɛstɚz) wɚ ər´ɛstɨd./

A crude syntactic analysis of the sentence is then performed based on locations of any orthographic commas, as well as the syntactic role of function words and verbs that are detected during the dictionary matching process. In the sample text just above, the verb *"were"* is detected and marked as the beginning of the verb phrase through use of the [ )] symbol. The end of a declarative sentence is indicated by the period symbol. The most important aspects of syntactic structure are the locations of clause boundaries, and the location of the boundary between the noun phrase and the verb phrase, although there are other syntactic factors that affect the rhythm and intonation of longer sentences. Liberal use of commas in text would help a great deal in formulating natural phrasing; their presence is generally a reliable cue, but unfortunately their absence does not indicate the absence of an intonational phrase boundary.

There is no semantic analysis in Klattalk or any other present-day text-to-speech system. Every sentence is spoken in a sort of semantically *"neutral"* way, i.e., without emphatic or contrastive stress, unless the user indicates an important word by placing the phonemic ["] symbol before it in the orthography.

Next, a phonemic representation is obtained for the words in the manner shown in Fig. 30. Each word is compared with entries in a small pronunciation dictionary. If no match is found, the word is broken into smaller pieces (morphemes) by attempting to remove common suffixes such as "-ed," "-ing," etc. It may be necessary to add a silent "e" or to reconstitute the "y" in order to recover the true form of the root, as in "biting = bite + ing." Then the remaining root is again compared with entries in the phonemic dictionary. If there is still no match, a set of letter-to-phoneme rules are invoked to predict the pronunciation. In this sentence, two words had affixes removed, five words/roots were found in the dictionary, and the remaining one was processed by letter-to-sound rules. No errors were made. The morpheme "protest" was found to have two alternative pronunciations in the dictionary, one with primary stress on the first syllable and the other with primary stress on the second syllable, but a selectional restriction associated with the "-er" suffix caused correct selection of the noun form.

TABLE III. Two-character representations for selected allophones in Klattalk.

Allophone	Two characters	Description
ɾ	DX	flap
ʔ	Q	glottal stop
tʼ	TX	glottalized t
ɚ	RX	postvocalic r
ɫ	LX	postvocalic l
iɚ	IR	vowel in "beer"
eɚ	ER	vowel in "bear"
ɑɚ	AR	vowel in "bar"
oɚ	OR	vowel in "boar"
uɚ	UR	vowel in "poor"

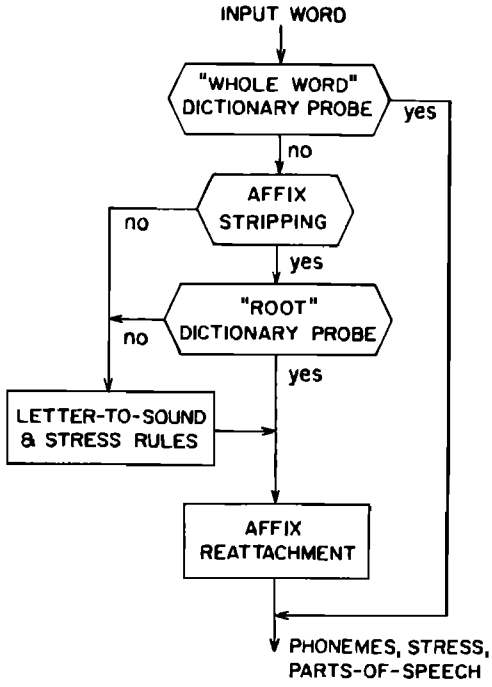

FIG. 30. The steps involved in converting an ASCII orthographic representation for a word into phonemes, stress, and parts-of-speech information. If the word or its base root is not in the dictionary, letter-to-sound rules guess at the proper pronunciation.

A part of the phonemic conversion process concerns the derivation of a stress pattern for the syllables of a word. Stress must be predicted if the word is not in the system vocabulary, or if the orthographic word is broken down into root plus affixes and an affix changes the stress pattern given for the root. The stress level of a syllable will be indicated by inserting a stress symbol just prior to the vowel in the phonemic representation. Absence of a stress symbol means that the syllable is unstressed.

A. Text formatting

A practical text-to-speech system has to be prepared to encounter words containing nonalphabetic characters, digit strings and unpronounceable ASCII characters. MITalk was one of the first systems to include algorithms for handling special cases such as how to speak digits in different formats (Allen *et al.*, 1979), e.g., "$35.61, 35.61, 2000, the year 1971, 10:15 p.m." This system also expanded many common abbreviations into full word equivalents. Commercial systems, which must be prepared to deal with more exotic material such as embedded escape sequences and other nonalphabetic characters, have adopted two general strategies. The Infovox SA-101 and the Prose-2000 provide the user with a set of logical switches which determine what to do with certain types of nonalphabetic strings. For example, " — " is translated to either "*dash*" or "*minus*" depending on the state of a switch. DECtalk, on the other hand, ignores escape characters, and usually spells out words containing nonalphabetic characters. The reasoning is that it is impossible to do the right thing in general, and the correct option for a particular application should be determined by a host computer. Even a simple strategy, such as interpreting a tab as an indicator of a new paragraph that should begin with a higher fundamental frequency, is not a safe assumption in arbitrary text; DECtalk therefore requires that a host computer insert a special "*new paragraph*" symbol in the text instead whenever tabs can be interpreted as new paragraphs. O'Malley *et al.* (1986) point out that many abbreviations are ambiguous, but can be disambiguated in particular applications. For example "N." is spoken as a letter in a name, as "North" in a street address, and as "New" in a state abbreviation, but these are easy fields to distinguish in a properly structured data base.

B. Letter-to-phoneme conversion

One issue in the preparation of rules and data structures for synthesis is how to best represent phonemes, allophones, stress, and syntactic symbols. Dictionaries generally do not agree on a standard representation, although the International Phonetic Association publishes one standard, and *The Journal of the Acoustical Society of America* employs a similar standard set of phonemic symbols that are used here in the examples. However, computers often require a representation that can be printed within the limitations of the ASCII character set. There is no agreement on either the set of phonetic symbols to be represented or the phonetic/alphabetic correspondences in this situation. The problem does not really require solution until such time as researchers wish to share data bases consisting of dictionaries or rules, and even then the most important issue is clear definition since computers are very good at symbol translation if they know what each symbol is intended to mean.

In my research, I have found it convenient to work with two different computer representations. One is case insensitive (upper case and lower case letters are equivalent) and requires two letters to represent vowels and some consonants. It is easy to type and easy to learn, so it is the way that words are input to Klattalk in phonemic form. The representation is nearly identical to the ARPAbet (Shoup, 1980). The second representation consists of a single ASCII character per phonetic symbol and so is an efficient way to store dictionaries and compare strings. Both representations can be parsed without the need for spaces between phonetic elements—in fact, "space" is the symbol used to indicate a word boundary. The two-character representation is defined and explained in Conroy *et al.* (1986, pp. 79–97), while the one-character set is described in Minow and Klatt (1983, Chap. 4). They are reprinted in Tables IV and V. The following are the only somewhat nonstandard symbols allowed in the abstract representation for a sentence: (1) there are two variants of schwa, /ə/ and /ɨ/, although the one to be used in any context is largely determined by the adjacent phonetic segments, (2) there is a separate symbol /yu/ for the more usual /y/ + /u^{w}/ because the fronting of /u^{w}/ in this environment would otherwise have to be done by a special rule, (3) there is a mapping of stressed /ɝ/ and unstressed /ɚ/ onto the single symbol /ɝ/, since this will cause no confusion and will make possible a slight saving in table space, (4) a silence phoneme is defined which is inserted by rule at cer-

TABLE IV. Two-character and one-character representations for phonemes in DECtalk.

Phoneme	Two characters	One character	Example
i	IY	i	beet
ɪ	IH	I	bit
e^y	EY	e	bait
ɛ	EH	E	bet
æ	AE	@	bat
ɑ	AA	a	pot
ɔ	AO	c	bought
ʌ	AH	^	but
o^w	OW	o	boat
ʊ	UH	U	book
u	UW	u	boot
ɝ	RR	R	Bert
$ɑ^y$	AY	A	bite
$ɔ^y$	OY	O	boy
$ɑ^w$	AW	W	bout
yu	YU	Y	Butte
ə	AX	x	about
ɨ	IX	|	nieces
p	P	p	pet
b	B	b	bet
t	T	t	tet
d	D	d	debt
k	K	k	kit
g	G	g	get
č	CH	C	Chet
ǰ	JH	J	jet
m	M	m	met
n	N	n	net
ŋ	NX	G	sang
f	F	f	fed
v	V	v	vet
θ	TH	T	thin
ð	DH	D	this
s	S	s	set
z	Z	z	zero
š	SH	S	shed
ž	ZH	Z	azure
w	W	w	wet
y	YX	y	yet
r	R	r	red
l	L	l	let
h	HX	h	head
n̩	EN	N	button
l̩	EL	L	bottle
_	_	_	silence "phoneme"

TABLE V. Two-character and one-character representations for stress and syntactic symbols in DECtalk.

Two characters	One character	Example
ˈ	ˈ	primary stress
ˌ	ˌ	secondary stress
"	"	emphatic stress
-	-	syllable boundary
*	*	morpheme boundary
#	#	compound noun
⟨space⟩	⟨space⟩	word boundary
(	(	begin preposition
)	)	begin verb phrase
,	,	clause boundary
.	.	end of sentence
?	?	question intonation
!	!	end of exclamation
+	+	new paragraph

tain syntactic boundaries, but can also be specified by the user, (5) one permitted special level of phrasal emphasis and two levels of lexical stress are introduced (plus the additional alternatives unstressed and reduced), (6) syllable boundary and morpheme boundary symbols are provided, but are usually not required since rules assign consonants to syllables correctly in most cases anyway, (7) the compound noun symbol is introduced in order to be able to force stress reduction in the second element of the compound, (8) a very limited inventory of syntactic symbols is provided, and (9) the new paragraph symbol is defined and used to realize prosodic marking of new paragraphs. A clear deficiency of the present symbol inventory is the lack of any ability to approximate non-English sounds in foreign words, although this limitation is shared by many English speaking individuals.

Historically, English and most other languages employing an alphabetical spelling representation began with a system that was close to the way the word was pronounced (Venezky, 1965, 1970; Chomsky and Halle, 1968; Henderson, 1982). Over time, pronunciation habits changed, sometimes dramatically, so that the spelling reflects more nearly an underlying historical antecedent of current pronunciation instead of the synchronic phonemes. Thus rules for pronunciation of English words depend on complex conventions involving, e.g., remote silent "e," the number of consonants following a vowel, the grouping together of special letter pairs, such as "ch" and "gh," which normally function like a single letter (Wijk, 1969), but not if in separate morphemes, etc. English has also borrowed words from other languages, so that Latin, French, German, and other patterns, somewhat Anglicized, are fairly common.

A selected survey of the literature on derivation of phonemes and stress from orthography is presented in Fig. 31 as a block diagram. Interconnections indicate how fundamental theoretical analyses of English have been incorporated in laboratory programs for text analysis and, finally, in commercial text-to-speech systems. As indicated in the figure, methods used in most commercial systems for deriving a phonemic representation of a word involve the use of letter-to-sound rules and an exceptions dictionary. An attractive alternative, as we will see, is to develop a large morpheme dictionary and try to decompose each input word into its constituent morphemes (where morphemes are the minimal meaningful subparts of words).

Several initial attempts to predict word pronunciation just from the spelling (Ainsworth, 1973; McIlroy, 1974; Hunnicutt, 1976; Elovitz *et al.*, 1976; Carlson and Granström, 1976) started from the assumption that a letter or letter pair could be converted to the appropriate phoneme if just the right amount of adjacent letter context was examined. Based on this view, a set of conversion rules was devised to take care of letter pairs such as "ch" and "ea," and

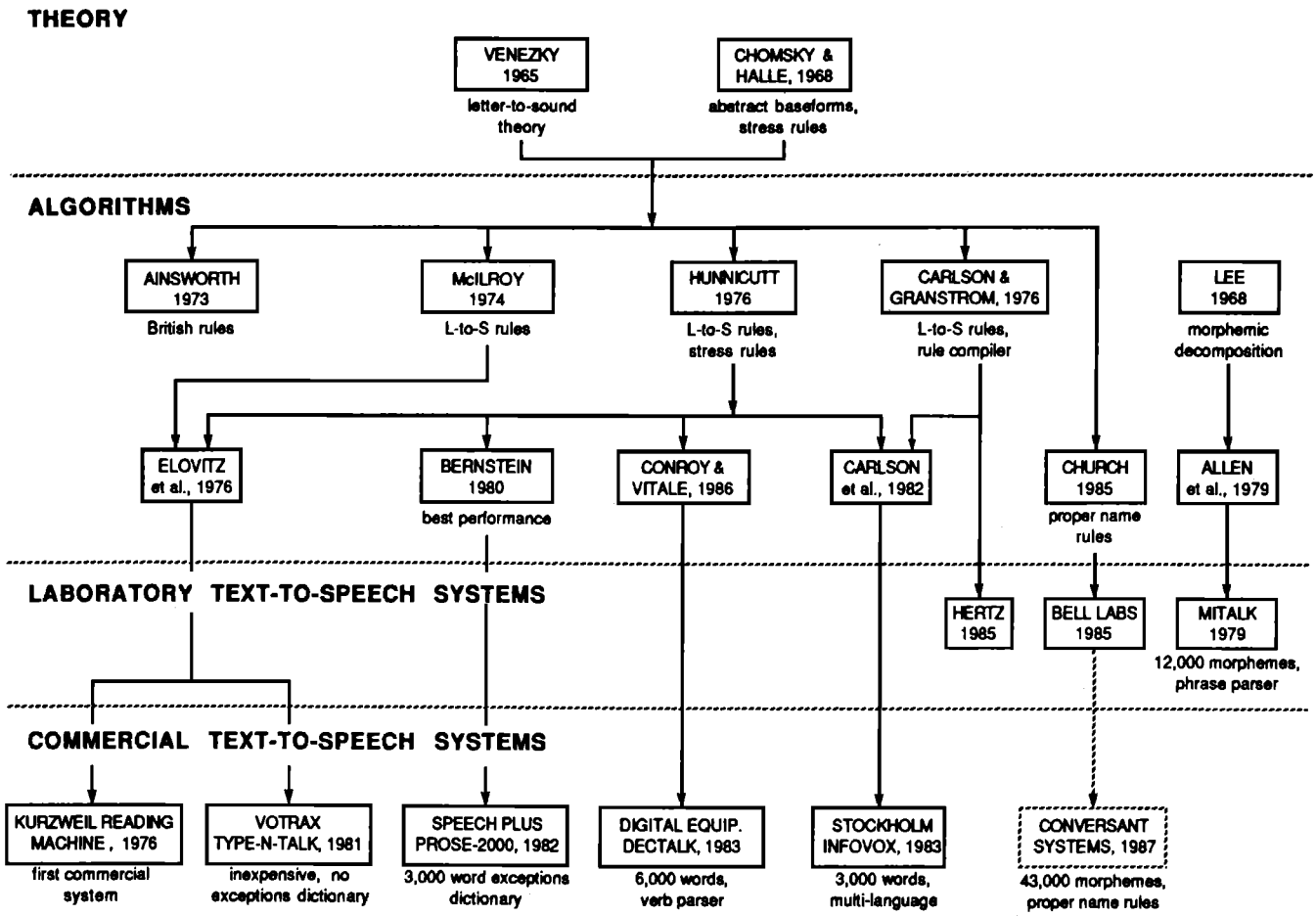

FIG. 31. Historical antecedents of the text-to-phoneme algorithms used in several laboratory and commercial text-to-speech systems.

then single letters were converted to phonemic form. For each letter, rules were ordered so that the first rules treated special cases of complex environmental specification, and the last case was always a default phonemic correspondence. For example, a rule might say that the letter A $\Rightarrow$ /e/ if followed by VE. The rule treats correctly words like "behave," but not "have." A slightly more complicated variant of the same idea was to convert consonants first (Hunnicutt, 1976). This permitted the phonemic representations of consonants to be used in the context specifications for the more difficult conversion of vowels. Systems of this kind may have more than 500 such rules for the interpretation of letter strings.

Several major problems were immediately apparent: (1) vowel conversion depended in part on stress pattern, (2) correct analysis often required detection of morpheme boundaries, and (3) letter contexts had structural properties such as VC vs VCC that one would rather refer to instead of enumerating all possible letter sequences. Before discussing how these and the next generation of spelling conversion programs dealt with these issues, we consider a novel approach to the problem that has received considerable attention in the artificial intelligence community.

The conversion of letters to phonemes might appear to be a pattern matching problem amenable to statistical learning strategies. For example, Sejnowski and Rosenberg (1986) considered the problem of creating a network, which they called NETtalk, that takes a seven-letter window as input and outputs the phoneme corresponding to the middle letter. A set of 120 "hidden" neuron-like threshold elements mediated between input neurons corresponding to 29 possible letters at each of seven positions and an output set of neurons representing about 40 phonemes and two degrees of stress. The weighting of input connections and output connections of the hidden units was initially random, but was adjusted through incremental training on a 20 000 word phonemic dictionary. When evaluated on the words of this training set, the network was correct for about 90% of the phonemes and stress patterns. In some sense, this is a surprisingly good result in that so much knowledge could be embedded in a moderate number of about 25 000 weights, but the performance is not nearly as accurate as that of a good set of letter-to-sound rules (performing without use of an exceptions dictionary, but with rules for recognizing common affixes). A typical knowledge-based rule system (Bernstein and Pisoni, 1980) is claimed to perform at about

85% correct at a word level (all phonemes correct and stress pattern correct) in a random sampling of a very large dictionary, which implies a phoneme correct rate of better than 97%.[10]

NETtalk is related to a letter-pattern learning program described earlier by Lucassen and Mercer (1984). They defined a set of features corresponding to "random" sets of letters, used the forward–backward algorithm of the IBM speech recognition strategy (analogous to incremental training) on a 50 000 word lexicon to find the best feature sets for predicting individual phonemes, and established a set of probabilities (analogous to weights) for a search tree recognition model, based again on a seven-letter input window. They obtained correct letter-to-phoneme correspondences for 94% of the letters in words in a random sample from a 5000 word office-correspondence lexicon. In terms of error rate, this is slightly better than NETtalk, especially considering that some fraction of the test words was probably not in the training set, but the Lucassen and Mercer approach still results in an inferior words-correct error rate compared with traditional rule systems. Even a very powerful statistical package cannot yet discover much of the underlying structure in a process as complex as natural language.

A proposal in the psychological literature related to these pattern learning programs is that readers learn the letter-to-phoneme conversion rules not as explicit rules, but by analogy with similar local letter patterns in words that they already know how to pronounce (Glushko, 1981; Dedina and Nusbaum, 1986). For example, a novel word might be compared with all words in the lexicon, and the word sharing the largest number of letters with the unknown word would get to determine the pronunciation of that local substring. Glushko showed that subjects were slower to pronounce pseudowords that would have two equally likely alternative pronunciations if this strategy were followed. A computer implementation of a slightly more complicated version of this strategy (taking into account frequency of occurrence of analogous words) agreed with one of the pronunciations furnished by human subjects 91% of the time when tested on 70 simple pseudowords (Dedina and Nusbaum, 1986), while DECtalk pronunciations agreed with the response from at least one of seven human subjects 97% of the time. Klatt and Shipman (1982) defined a way in which the substring comparison strategy might be performed optimally and rapidly, one letter at a time, by creating a moderate-sized decision tree. They examined the performance when a 20 000 word phonemic dictionary was divided in half randomly such that the first half was used to create the tree, and the second half used to test it. The error rate for individual letters was 7%, which is not bad considering that test and training data were different, but this performance is still not nearly good enough to compete with conventional rule systems. Consonantal letters were found to be quite regular and amenable to translation with low error rates by this approach. However, the five vowel letters and "Y" accounted for four-fifths of the errors. In summary, given the attention that NETtalk and other neuron-like devices have received recently, it is disturbing that NETtalk does not learn training set data perfectly, appears to make generalizations suboptimally, and has an overall performance that is not acceptable for a practical system. Furthermore, it is unlikely that larger training lexicons would converge to a more acceptable performance. Aside from limitations imposed by the network model, problems inherent in all these approaches are (1) the considerable extent of letter context that can influence stress patterns in a long word (and hence affect vowel quality in words like "photograph/photography"), (2) the confusion caused by some letter pairs, like CH, which function as a single letter in a deep sense, and thus misalign any relevant letters occurring further from the vowel, and (3) the difficulty of dealing with compound words (such as "houseboat" with its silent "e"), i.e., compounds act as if a space were hidden between two of the letters inside the word. The necessity of morphemic analysis is supported by data indicating that good spellers look for morphemes inside letter strings (Fischer *et al.*, 1985), whereas to date these learning models seek regularities in letter patterns without recourse to a lexicon of any sort. On the other hand, efforts to find clear psychological evidence for morphological analysis of complex related forms (as opposed to rote learning of each) for word pairs such as "heal-/health," "original/originality," "magic/magician," and "sign/signal" have generally failed (Carlisle, 1985).

1. Prediction of lexical stress from orthography

The Hunnicutt (1976) rule system included the improved version of Chomsky–Halle stress rules (Halle and Keyser, 1971) consisting of eight general rules, the most well-known of which are the main and alternating stress rules for predicting which syllable receives primary stress as a function of the "strong/weak" syllable pattern of the word. Also included were rules for decomposing words by stripping off affixes to recover the root. About 15 different prefixes and 50 suffixes were detected. Grammatical constraints were invoked to prevent incompatible suffix sequences from being removed. Orthographic features permitted rules to refer to concepts such as "true consonant" and "vowel-like letter." While the best performing algorithm of its time, this system was completely correct for only about 65% of a random selection of words (Hunnicutt, 1980).[11] A good fraction of the errors made by this letter-to-phoneme system were stress errors. In fact, Bernstein and Nessly (1981) showed that a much simpler set of stress rules described by Hill and Nessly (1973) performed about as well as the Chomsky–Halle implementation. More recent high-performance letter-to-phoneme rule systems (Bernstein and Nessly, 1981; Hunnicutt, 1980; Hertz, 1982; Carlson *et al.*, 1982a; Church, 1985, Conroy and Vitale, 1986) include improved attempts at morphemic decomposition and stress prediction. Stress assignment is perhaps the weakest link in all systems because an incorrect stress pattern, while perceptually disruptive in and of itself, usually also triggers mis-selection of vowel qualities. The newer systems not only base stress assignment on factors such as morphological structure and the distinction between strong and weak syllables (Chomsky and Halle, 1968), but also on presumed part of speech, and in some cases, etymology (for a good review, see Church, 1985). The importance of syntactic categorization is sug-

gested by statistics indicating that over 90% of bisyllabic nouns have stress on the first syllable, while only about 15% of bisyllabic verbs are stressed on the first syllable (Francis and Kučera, 1982).

One issue faced by designers of systems is which to do first, stress prediction or phoneme prediction. Another issue is whether to essentially work forward or backward through the letter string for a word. While no system has to go only left to right, or completely settle stress prediction prior to phonemic analysis, there seem to be clear advantages to working backwards through the letter string, and to having stress information prior to making vowel decisions (Bernstein and Nessly, 1981).

2. Exceptions to the rules

When evaluating a set of letter-to-phoneme rules, it is easy to make up lists of words that fail to be pronounced properly. Systematic comparison of the rules against a list of frequent words can produce a dictionary of exceptions that, if added to the system, will make overall pronunciation performance much better than for a system that only uses rules. The utility of a small exceptions dictionary can be appreciated by observing the ability of a small number of most frequent words to account for a given fraction of words in running text (Hunnicutt, 1980). The data are reproduced in Fig. 32. They indicate that a small number of words, about 200, are required to cover half the words occurring in a random text. With a dictionary of 2000 words, over 70% of the words in text will be matched and not have to go through letter-to-sound rules. However, the law of diminishing returns begins to take over shortly after this point—if one extrapolates from the slope of the curve prior to 10 000 words,[12] as indicated by the dashed line in Fig. 32, it appears that to go from 90% to 93% coverage would require about an additional 60 000 words!

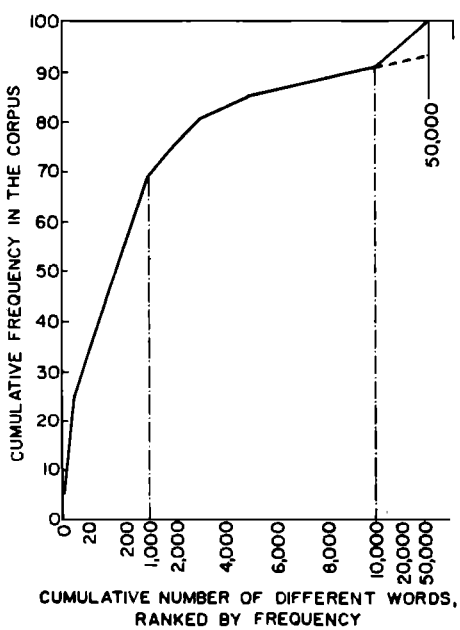

FIG. 32. A one million word corpus (Kučera and Francis, 1967), containing about 50 000 different words, can be used to estimate the number of entries in a lexicon necessary to match a given percent of words in a new text, after Hunnicutt (1980).

Elovitz *et al.* (1976) and Hertz (1982) embed lists of exceptions inside the letter-to-sound rules of their systems (such as the observation that the letter "f" is pronounced with a voiceless/f/ phoneme in all words except "of") so as to ensure getting common words correct, whereas others tend to segregate out exceptions as a separate dictionary. The best performance for a rule system without exceptions dictionary, better than 85% correct when tested on a random sample from a large dictionary, has been obtained by the Bernstein rules that are a part of the Speech Plus, Inc. Prose-2000 (Groner *et al.*, 1982). Bernstein argues that it is possible to design a letter-to-sound algorithm with a very simple structure—consisting of one right-to-left pass through the letters, starting inside all stress-neutral suffixes.

A moderate-sized exceptions dictionary can hide the deficiencies of a weak set of letter-to-sound rules, but at a high cost in terms of storage requirements. Based on data shown in Fig. 32, Hunnicutt (1980) showed that the size of an exceptions dictionary required to get a target fraction of input words pronounced correctly in a typical running text is a strong function of letter-to-sound rule performance. For example, the 3000-word exceptions dictionary in the Speech Plus Prose-2000, coupled with rules that are correct 85% of the time, results in an overall system performance of better than 97% correct (only 1 word in 33 in a typical text contains a noticeable phoneme or stress error). On the other hand, the first version of DECtalk, employing the Hunnicutt (1980) rules with 65% accuracy and a larger 6000-word exceptions dictionary, barely reached 95% correct (1 error every 20 words). Independent confirmation of this accuracy comparison comes from Huggins *et al.* (1986), who examined over 1600 low-frequency polysyllabic words and found phonemic mispronunciations in 8.3% for the new Speech Plus Calltext system, compared with 12.9% errors for Version 1 of DECtalk. The current DECtalk, Version 3.0, uses a new letter-to-sound rule system (Conroy and Vitale, 1986) to achieve performance of fewer than 6% errors for this data set, according to my evaluation.

In the future, it is expected that morpheme-based algorithms (see below) will replace exceptions dictionaries in commercial systems because the cost of memory is such that the added performance is well worth the expense. Similarly, special algorithms for pronunciation of names are likely to be incorporated in commercial systems in the near future. Special purpose vocabularies, such as a dictionary of medical terms, will probably also become available in response to market pressures.

3. Morphemic decomposition

Problems with the pronunciation of compounds such as the "*th*" in "*hothouse*" and the silent "*e*" in "*houseboat*" led Lee (1969) to attempt to break each word into morphemes, the minimal meaningful unit of language (see, e.g., Bloomfield, 1933, Chaps. 10, 13–14). Using a dictionary of about 3000 morphemes, Lee was able to split a word such as "*houseboats*" into "*house*" plus "*boat*" plus the plural "*-s*," and to retrieve from storage or predict the pronunciation of each piece. Lee developed techniques for recovering the proper base form after an affix was removed. The three most common problems, which could be handled correctly most

of the time using morphological decomposition, involved situations when the surface form did not contain a silent "e" (choking→choke + ing), there had been consonant doubling (omitted→omit + ed) or a final "y" had been modified (cities→city + s). Jonathan Allen and Deborah Finkel extended these techniques by increasing the morpheme dictionary to 12 000 items (Allen *et al.*, 1979; Allen *et al.*, 1987). Morphemes were selected by interactive examination of the approximately 50 000 unique words in the Brown corpus, a sampling of one million words of text (Kučera and Francis, 1967).[13]

Allen *et al.* (1979) also developed rules for handling cases where a word has multiple parses (e.g., "*scarcity*" = "*scarce* + *ity*" or "*scar* + *city*"). One rule, illustrated by this example, is that affixing is more likely than compounding. None of these guidelines is absolute, so in comparing two alternative morphemic decompositions, the authors invoked a set of heuristic scoring procedures whereby a given morphemic division incurs a scoring penalty depending on what has happened so far. This scoring algorithm picks the correct decomposition for "formally" from among the set ⟨form + all + y, for + mall + y, form + ally, form + al + ly⟩. If, after all of this computation, the word was found to be an exception to the parsing heuristics (e.g., "*been*" not pronounced as "*be*" + "*-en*"), the whole word was added to the morpheme lexicon in unparsed form.[14] An alternative method for dealing with inflectional suffixes, derivational affixes, and compounding is discussed in Church (1985, p. 251).

Some morphemes are pronounced differently depending on the stress pattern of the word and the nature of the other morphemes present (note the second "*o*" of "*photo*" is realized phonemically as /o,ə,ɑ/ in "*photo*," "*photograph*," "*photography*," respectively). The MITalk group developed rules to handle some of these cases, and simply added whole multimorphemic words to the lexicon if the rule was too complex or not sufficiently productive. The morpheme decomposition algorithm is able to parse about 98% of the words in a typical text, and should have greater accuracy than letter-to-phoneme rules. The exact accuracy of the MITalk morpheme decomposition algorithm was never measured, although a cursory glance at a three-paragraph text (Allen *et al.*, 1987, pp. 89–92) indicates a few (easily correctable) errors and a words-correct rate of only about 95%.

One of the advantages of a morpheme lexicon, aside from an ability to divide compound words properly, is that a set of 12 000 morphemes can represent well over 100 000 English words. Thus a very large vocabulary is achieved at moderate storage cost. However, the greatest advantage of the morpheme lexicon may turn out to be its ability to specify parts of speech information to a syntactic analyzer in order to improve the prosody of sentences, see below.

Recent work at Bell Laboratories (Coker, 1985) has extended this approach by augmenting the morpheme lexicon to 43 000 morphemes, and adding to the rules for suffix and prefix analysis and stress reassignment for the stress-shifting suffixes. The algorithm and morpheme lexicon occupy about 900 kbytes on a developmental real-time text-to-speech board (Olive and Liberman, 1985).

4. Proper names

Proper names are a special problem because the rules for their pronunciation often depend on which language is assumed as the underlying origin of the spelling (Liberman, 1979). The commercial system that performs best at pronouncing proper names, the newest Speech Plus Calltext board, still has an error rate of about 20% in its rule component when confronted with random proper names (Wright *et al.*, 1986). Church (1985) has recently proposed a solution to this problem that involves statistics on the frequency of occurrence of three-letter sequences in each of several languages. The first step is to use these statistics to estimate the language family of the unknown word. For words of moderate length, he finds that frequently one or another letter triple in the word essentially rules out all but the correct language. The second step is to apply stress and letter-to-phoneme rules for the language in question. Performance is claimed to be far superior to that of any system restricted to a single set of rules for all proper names. The importance of doing proper names by rule is brought out by statistical analyses showing that large name dictionaries do not solve the problem. An exceptions dictionary containing 2000 proper names will cover about 50% of the names in a random telephone directory, and 6000 proper names will cover about 60%. However, adding to the exceptions dictionary beyond 6000 names is essentially fruitless in that one is unable to get beyond an asymptote of about 62% of the names in one telephone directory, no matter how many names are obtained from another directory (Church, 1985).

C. Syntactic analysis

Imposition of an appropriate prosodic contour on a sentence requires at least a partial syntactic analysis. Furthermore, some pronunciation ambiguities can be resolved from syntactic information. For example, there are more than 50 noun/verb ambiguous words such as "*permit*" that are pronounced with stress on the first syllable if a noun, and with stress on the second syllable if a verb (see Appendix D in Conroy *et al.*, 1986). The only way to pronounce these words correctly is to figure out the syntactic structure of an input sentence, including the location of the verbs. Proper phrasing of moderately long clauses also requires knowledge of the locations of phrase boundaries. Thus it would be highly desirable to include a parser in a text-to-speech system.

While powerful parsing strategies exist (see, e.g., Woods, 1970; Aho and Ullman, 1972; Marcus, 1980; Kaplan and Bresnan, 1982), they tend to produce many alternative parses, even for sentences that seem simple and unambiguous. For example, "Time flies like an arrow" is multiply ambiguous at a *syntactic* level; a syntactic analysis system would require an immense store of world knowledge (semantics/pragmatics) to behave as we do and focus immediately on the only sensible structural interpretation of the sentence. Allen (1976) foresaw this problem and restricted himself to the goal of selecting the most probable local phrase parse of an arbitrary English sentence. Using the morpheme decomposition algorithm just described, he and Calvin Drake were able to obtain reasonably accurate

part-of-speech alternatives for most words of the sentence from the morpheme decomposition routine, and assumed tentatively that all unanalyzable words were nouns. The syntactic analysis proceeded left-to-right, attempting to add as many words as possible to each phrasal constituent. A backup algorithm suggested by Lorinda Cherry at Bell Laboratories sought possible verbs if it turned out that this process failed to recover a verb, as would be the case when a noun/verb ambiguity like "*permit*" was present in a sentence such as "Police permit mopeds." While the performance of this parser was never extensively tested, examination of some sample texts (Allen *et al.*, 1987, pp. 89–92) suggests that it works reasonably well, but produces several inappropriate pauses and pseudopauses at falsely detected boundaries.

If a parts-of-speech categorization is not available for most words, the simplest parsing strategy would be to use function words such as prepositions, conjunctions, and articles to find obvious phrase boundaries, leaving the remaining boundaries undetected. This is the strategy employed in the Prose-2000 and in the Infovox SA-101. The Votrax Type-n-Talk appears to use only punctuation marks as parsing cues.

DECtalk employs not only function words, but also a moderate-sized dictionary of verbs that unambiguously indicates the beginning of a verb phrase (Klatt, 1975a). Detection of the beginning of a verb phrase in a long clause permits DECtalk to break the intonation contour into two rise–fall "*hat-pattern*" units that help the listener parse the sentence. However, it is better to miss a noun-phrase/verb-phrase boundary than to insert prosodic boundary gestures (fall–rise intonation contour and lengthening of a phrase-final syllable) at locations where they do not belong. In an earlier experimental system that assumed that any word that could be a verb was a verb, listeners were distracted and often confused by extra prosodic boundaries, while the absence of a prosodic gesture just sounded like the speaker was talking too fast. DECtalk also provides a simple mechanism for a user to indicate a phrase boundary when one is missed—the [)] symbol can be inserted between the words in question. DECtalk does not try to disambiguate noun/verb ambiguities; the most frequent pronunciation is given unless the user requests the second most frequent pronunciation by attaching a special symbol to the front of the orthography.

DECtalk and other text-to-speech systems make a large number of syntactic errors that lead to noticeable misphrasings. In the future, syntactic routines will be expected to provide better detection of the following:

- phrasal constituency—particularly the locations of left-branching constituents and non-adjacent sister constituents that should probably be marked by prosodic gestures,
- internal structure and compounding relations within long noun/adjective strings,
- when to "pop" from an embedded clause that is not terminated by a comma,
- how to determine the nature of conjoined units on either side of a conjunction so as to be able to insert a syntactic break when appropriate,
- syntactic deletion sites where some sort of prosodic gesture should be synthesized to indicate the location of the missing material (Cooper *et al.*, 1978),
- how to detect tags and parenthetical material such as "This is the answer, he told us," that are usually said in a noninflected way,
- resolution of part-of-speech ambiguity, for (1) words that can be either an unstressed preposition or a stressed verbal particle such as "*on*" in "*He takes on hard jobs*," (2) instances where "*that*" is functioning as a (stressed) demonstrative, e.g., "*I know (that) THAT book is red*" rather than as an unstressed clause introducer, as in "*I know that books are red*," and (3) instances of compounds that are pronounced with reduced stress on the second word, such as "*He lived in Baker House* (this is largely a lexical/semantics problem).

D. Semantic analysis

Semantic and pragmatic knowledge is needed to disambiguate sentences like the ones the New Yorker is fond of reprinting. For example, in a sentence such as "*She hit the old man with the umbrella*," there may be a pseudopause (a slowing down of speaking rate and a fall–rise in pitch) between the words "*man*" and "*with*" if the woman held the umbrella, but not if the old man did. Similarly, a "*rocking chair*" will have the word "*chair*" destressed if the combination of adjective and noun has been associated by frequent use into a single compound-noun entity. Emphasis or contrastive stress may be applied to an important word depending on the meaning: "*The OLD man sat in a rocker*" (not the younger man). Finally, words that have lost their importance in a dialog, either because of prior occurrence of the word or by anaphoric reference, should be destressed.

No text-to-speech system is capable of dealing automatically with any of these issues. DECtalk employs the simplest possible solution by providing the user with an input inventory of symbols to facilitate user specification of the locations of missing pseudopauses (the [)] symbol), unmarked compound words (spell as "*rocking-chair*"), and emphasis (precede the emphasized word by an emphasis symbol ["]).

It is possible to think of applications where the computer is not simply attempting to speak ASCII text, but may know a great deal about the meaning of the message, perhaps having formulated the text from a deep-structure semantic representation in, e.g., a data base information retrieval application (Young and Fallside, 1979). In such cases, one would want to take advantage of the ability to mark for emphasis important words when forming the input to the text-to-speech system. Hirshberg and Pierrehumbert (1986) provide an excellent review of the factors influencing the intonational structuring of discourse.

In the future, systems that have available parts-of-speech information from a large morpheme lexicon can be expected to develop better syntactic analysis routines that are particularly suited to the problems of text synthesis. Perhaps computer science efforts to produce expert systems will

lead to advances in semantic representation that can be adapted to text synthesis as well.

III. HARDWARE IMPLEMENTATION

A laboratory text-to-speech system, or a development system, is best implemented on a large general-purpose digital computer. The flexibility and nearly unlimited computational resources outweigh disadvantages of non-real-time output. However, practical commercial systems must realize real-time operation at a reasonable cost/performance tradeoff, while simultaneously providing additional features such as a flexible user interface and telephonics for many commercial applications. Solutions may require specially designed chip sets (Gagnon, 1978; Goldhor and Lund, 1983) or circuit boards containing off-the-shelf components rich in computer power and memory (Groner *et al.*, 1982; Bruckert *et al.*, 1983).

One important design consideration is the sampling rate and resultant high-frequency cutoff of the output speech. Since many business applications require the telephone, some systems limit the frequency response to that of telephone bandwidth—3.4 kHz, or the 4.0-kHz limit imposed by the 8-kHz sampling rate of standard codec digital transmission of speech (Groner, 1982; Olive and Liberman, 1985). DECtalk, on the other hand, produces information at frequencies up to 5 kHz in order to maximize intelligibility over a loudspeaker in, e.g., handicapped applications, such as a reading machine for the blind.

My own experiences may help illustrate hardware issues. In order to transform the Klattalk software into a real-time device, it was necessary for me to find a commercial partner with the appropriate skills and deep pockets. Fortunately, Digital Equipment Corporation was willing to underwrite the development costs. We signed a license agreement in 1982 (Klatt, 1987), and a product, DECtalk, was announced some 18 months later (Bruckert *et al.*, 1983).

The DECtalk hardware, Fig. 33, was capable of implementing the complete existing Klattalk software; no engineering compromises were necessary. Software added by Digital engineers controlled the user interface to a host computer. Host computer commands were defined to permit initiation or reception of telephone calls, and to permit the host to suddenly halt speaking, or to monitor the instant when a particular word in a sentence has been spoken.

The hardware shown in Fig. 33 includes (1) a Motorola MC68000 general purpose digital computer that processes text corresponding to one clause at a time, producing a set of synthesizer control parameters every 6.4 ms, and (2) a Texas Instruments TMS-32010 signal processing chip that converts control parameters to difference equation constants, and simulates the digital formant synthesizer in order to produce 10 000 12-bit waveform samples per second. Memory requirements are modest. The 6000-word exceptions dictionary places the greatest demands on memory; it occupies about half of the read-only memory shown in the figure. DECtalk can be controlled by any computer or by an ordinary computer terminal since the communication link is via a standard RS-232 port.

The only disappointment was that the price of the original DECtalk system turned out to be about four times our early estimate of $1000, and this placed the device outside the reach of many potential handicapped users. A recent redesign of the main DECtalk board to contain less "integrated circuit glue" has resulted in the DECtalk 3.0 system that is improved in several performance areas and is less expensive to manufacture, so there is still hope that an acceptable price might be achieved. Board size, about $8 \times 10 \times 0.7$-in. sans power and loudspeaker, is now satisfactory for portability, but lower power consumption is a goal that will have to be met in the future.

Today's technology is such that, I am told, it would be possible to put the entire text-to-speech algorithm on a single wafer-sized integrated circuit chip. However, this is not likely to happen until the demand is sufficient to justify chip design costs. Instead, it appears that future versions of the hardware may move toward greater flexibility by replacing all of the read-only memory by RAM that can be down loaded with new code as algorithms are improved.

IV. PERCEPTUAL EVALUATION OF TEXT-TO-SPEECH SYSTEMS

Text-to-speech systems can be evaluated and compared with respect to intelligibility, naturalness, and suitability for particular applications. One can measure the intelligibility of individual phonemes, words, or words in sentence context, and one can even estimate listening comprehension and

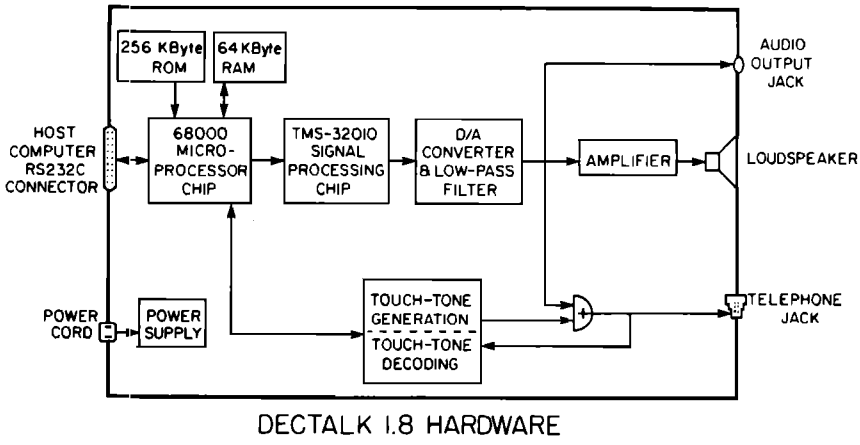

FIG. 33. Electronic hardware used in the first DECtalk implementation of Klattalk.

cognitive load, see Table VI. There are only a few studies that have attempted quantitative evaluations of text-to-speech systems to date; much of the data on the capabilities and limitations of the current technology comes from work performed at Indiana University by David Pisoni and his colleagues (Pisoni *et al.*, 1985).

A. Intelligibility of isolated words

The measurement of intelligibility can be performed in many different ways. Since consonants have been more difficult to synthesize than vowels, the modified rhyme test (House *et al.*, 1965) is often used, in which the listener selects among six familiar words that differ only by an initial consonant or a final consonant. This is not a very severe test of system performance since the response alternatives may exclude a confusion that would be made if a blank answer sheet were used, but the test does facilitate rapid presentation to naive subjects and automatic scoring of answer sheets.[15] If possible, an open response, including perhaps a rating of goodness of each item, should be used with such a test in order to better determine systematic error patterns and deficiencies, especially if there are relatively few errors.

Logan *et al.* (1986) evaluated the intelligibility of eight text-to-speech systems by presenting listeners with a recording of the modified rhyme test words. The results are summarized in Table VII. Also included are comparable data obtained earlier with the Haskins text-to-speech system (Cooper *et al.*, 1984). Systems are rank ordered according to performance. When percent correct is fairly high, a good way to compare systems is to use percent error (simply 100 minus percent correct) because relative changes in percent error better reflect the difficulty of comprehension and the difficulty of making improvements. The frequency of occurrence of perceptual errors in running text is approximated by the reciprocal of the percent error values given in the table.

TABLE VI. Techniques for evaluating text-to-speech systems.

INTELLIGIBILITY:	Diagnostic rhyme test (Fairbanks, 1958; Voiers, 1983)
	Modified rhyme test (House *et al.*, 1965)
	Open response rhyme test (Pisoni *et al.*, 1985)
	MRT in noise (Nixon *et al.*, 1985)
	CNC word list (Lehiste and Peterson, 1959)
	CVC nonsense words (Dubno and Levitt, 1981; Pols and Olive, 1983)
	CID W-22 word list (Hirsh *et al.*, 1952)
	Goodness ratings for words (Wright, Altom, and Olive, 1986)
	CID sentences (Erber, 1979)
	Harvard sentences (Egan, 1948)
	SPIN test (Nakatani and Dukes, 1973; Kalikow *et al.*, 1977)
	Haskins anomalous sentences (Nye and Gaitenby, 1974)
COMPREHENSION:	Reading/listening comprehension (Pisoni and Hunnicutt, 1980)
	Sentence verification (Manous *et al.*, 1985)
NATURALNESS:	Paired comparisons (IEEE, 1969; Logan and Pisoni, 1986)
	Subjective ratings (Nusbaum *et al.*, 1984)

TABLE VII. Performance of selected text-to-speech systems with respect to CVC intelligibility using the modified rhyme test, closed response, after Logan *et al.* (1986) and Cooper *et al.* (1984).

Device	% correct	% error
Type-n-Talk	73	27
Infovox	88	12
MITalk-79	93	7
Prose-2000 3.0	94	6
DECtalk 1.8	97	3
Natural speech	99	1
Haskins system	93	7
Natural speech	98	2

Looked at in this way, the expected rate of perceptual errors for DECtalk is about (100%/3%), or one segmental misperception about every 33 syllables of text. The error rate for the Prose-2000 is about twice that of DECtalk, while it appears that Type-n-Talk is seriously flawed (see also Cochran, 1986).

When Logan *et al.* (1986) ran the same vocabulary used in the modified rhyme test, but with open response, the error rate went up quite a bit—typically 3 to 4 times the closed-response error rate—but the relative rankings of systems did not change. Open response, however, had the advantage that systematic error tendencies could be detected and (hopefully) corrected. For example, DECtalk 1.8 had a problem with nasals adjacent to high front vowels—a problem that was then corrected in DECtalk 3.0. The test used is perhaps not ideal for detection of all likely consonantal confusions because the words are not particularly well balanced phonetically, and there are no consonant clusters or unstressed syllables. Other word lists address some of these deficiencies (Lehiste and Peterson, 1959; Nusbaum *et al.*, 1984), but there is a clear need for better diagnostic instruments in the evaluation of text-to-speech systems.

The intelligibility of several linear prediction based systems has been studied by Pols and Olive (1983). They presented consonant–vowel–consonant (CVC) nonsense syllables to high school students after a brief introduction to phonemic representations. The syllables were either (1) natural speech digitized at 10 000 12-bit samples/s, (2) 10-pole linear-prediction coded versions of these syllables, or (3) syllables synthesized using the Olive (1977) LP diphone concatenation scheme. The results are shown in Table VIII. This is a very difficult task for naive unpracticed subjects, as indicated by the relatively low 93% phoneme recognition performance for natural speech.[16] Two points of interest are that (1) linear prediction coded speech can suffer a serious reduction in intelligibility, even when there is no effort to

TABLE VIII. Consonant intelligibility in nonsense syllables encoded in various ways (Pols and Olive, 1979).

Condition	% correct	Typical errors
OLIVE (1977) DIPHONE SYNTHESIS	66	voicing, nasality
LPC-10, no quantization	86	b-v-ð,m-n-ŋ
DIGITIZED NATURAL, 5 kHz, 12 bit	93	f-θ,v-ð

save bits by quantizing the representation, and thus (2) linear-prediction coded speech often gives listeners more favorable impressions of intelligibility and naturalness than are warranted by objective measures.

Based on this critical evaluation, Olive went on to select new versions of his diphone inventory, also hand-correcting pitch errors, and retested diphone intelligibility iteratively until the most recent system exceeds the intelligibility of LP-10. Part of the intelligibility increase may be attributed to the use of multipulse linear prediction (Atal and Remde, 1982; Olive and Liberman, 1985), which makes possible the detailed modeling of bursts of noise and other syllable-onset events.[17]

Wright *et al.* (1986) discovered that it is possible to detect deficiencies in segmental synthesis even when intelligibility is relatively high, simply by asking subjects to rate the subjective goodness of words. Naive listeners hear a word and then see a visual presentation of the word, at which point they are asked to rate goodness. If the goodness rating is low, the computer asks additional questions about the location and type of specific defects.

In an effort to find a maximally sensitive test for comparing phoneme intelligibility of various systems, Nixon *et al.* (1985) added controlled amounts of background noise to synthesized or vocoded MRT word lists, and measured intelligibility as a function of signal-to-noise level. They found that an unidentified "high-performance" text-to-speech system was about six percentage points worse than natural speech over a wide range of S/N ratios. Stated in another way, under adverse S/N conditions, the synthetic speech had to have a 5-dB boost in S/N ratio to be as intelligible as natural speech.[18] Of perhaps greater interest are comparative figures from the Nixon study for 2.4-kbit government-standard LPC-10, and 9600-bit CVSD, both of which performed much worse than the synthetic speech produced by this text-to-speech system—both being about 40% less intelligible than natural speech at high and low S/N ratios. These rather surprising results suggest limits to the utility of low-bit-rate encoded speech, and suggest that, at least for some applications, text-to-speech systems already offer superior communicative performance.

B. Intelligibility of words in sentences

In comparison with words spoken in isolation, words in sentences undergo significant coarticulation across word boundaries, phonetic simplifications, reduction of unstressed syllables, and prosodic modifications that, among other things, shorten nonfinal syllables and modify the fundamental frequency contour. In order to evaluate the ability of text-to-speech systems to realize these transformations, tests of word intelligibility in sentence frames have been devised. The easiest materials, consisting of simple short predictable sentences known as the CID sentences (Erber, 1979), have been used primarily to evaluate abilities of the hearing impaired. Another sentence list was devised to measure speech intelligibility in noise (Egan, 1948). This list, known as the Harvard sentences, is often employed today, in spite of its meager syntactic variation and minimal use of words with more than two syllables, simply because no better lists have been proposed and calibrated. Pisoni *et al.* (1985) employed a subset of the Harvard Sentences and measured the intelligibility of each content word. The results are presented in Table IX. The same rank order of systems holds as was obtained for isolated words. Also shown in the table are data from a Haskins anomalous sentence test (Nye and Gaitenby, 1974), consisting of nonsensical word strings that were syntactically acceptable—of the form "The (adjective) (noun) (verb) the (noun)," e.g., "The old farm cost the blood." Again, system rank ordering is the same, but differences between systems are somewhat greater, suggesting that this is a more sensitive test.

The performance of the Haskins system, as evaluated by Ingemann (1978) and reported by Cooper *et al.* (1984) is also shown in the table. The poorer general performance of subjects in the Ingemann study on the natural speech control may imply that the scores should be boosted slightly before comparison with the systems listed above in the table.

Chial (1985) used the SPIN (speech in noise) test developed by Kalikow *et al.* (1977) and calibrated by Bilger *et al.* (1984) to evaluate the relative performance of several of the less expensive text-to-speech systems. Subjects had to identify the last word in sentences presented in a background babble of several competing voices. Included were the Echo II, the Votrax Type-n-Talk that incorporates the SC-01 synthesis-by-rule chip, and the Votrax Personal Speech System, which uses a new version of the chip, the SC-01A. Results, shown in Table X, indicate that the new chip has improved intelligibility over the SC-01, up from 40% to 65% words correct as measured at 0-dB signal-to-babble level. However, performance with natural speech at this signal-to-babble level is typically about 91% correct (Chial, 1985; Bilger *et al.*, 1984), so one must conclude that these inexpensive devices are still very limited in intelligibility.

C. Reading comprehension

Since synthetic speech is less intelligible than natural speech, what happens when one tries to understand long paragraphs? Do listeners miss important information? Is a listener so preoccupied with decoding individual words that the message is quickly forgotten? In an attempt to answer these questions, Pisoni and Hunnicutt (1980) included a standard reading comprehension task in their evaluations. Half the subjects read the paragraphs by eye, while the other half listened to a text-to-speech system. In a later experiment, comparison was made with a human voice reading the

TABLE IX. Performance of selected text-to-speech systems with respect to word intelligibility in Harvard test sentences and Haskins anomalous sentences, after Pisoni *et al.* (1985) and Cooper *et al.* (1984).

Device	Meaningful % correct	Anomalous % correct
Prose-2000	84	65
MITalk-79	93	79
DECtalk	95	87
Natural speech	99	98
Haskins system		78
Natural speech		95

TABLE X. Performance of inexpensive text-to-speech systems with respect to word intelligibility in the SPIN (speech in noise) test sentences, after Chial, 1985.

Device	% correct
ECHO-II	18
VOTRAX TYPE-N-TALK	40
VOTRAX PERSONAL SPEECH SYSTEM	65
Natural Speech	91

paragraphs. Results of answering multiple-choice questions about the content of the paragraphs are shown in Table XI. The text-to-speech systems performed about equally well, suggesting that the test is not sensitive enough to compare systems, and that the limit on performance is the memory capacity of these college students rather than the difficulty of comprehending synthetic speech. Pisoni also observed that subjects typically got better on the second half of the test when listening to synthetic speech, even though there was no feedback of correct answers.[19] On the second half, listening subjects performed about as well as the readers.

One might conclude that current text-to-speech systems produce quite satisfactory speech since there is no measurable decrement in listening comprehension after a familiarization period. Thus synthetic speech should be a viable method of presenting information over an auditory channel in most applications. Such conclusions are perhaps premature because (1) similar experiments have not been performed over the telephone, or with less-educated subjects, and (2) multiple-choice tests and recall measures may not be sensitive enough to reveal differences in perceptual processing between natural and synthetic speech. Pisoni (1982) used a reaction-time experiment to show that listeners do indeed devote somewhat more time to speech perception when exposed to synthetic speech as compared with natural speech, and Manous *et al.* (1985) measured a decrement in accuracy and speed of response for text-to-speech systems versus natural speech using a more sensitive comprehension test in which listeners had to immediately respond "true" or "false" to each sentence they heard. The capacity of short-term memory for earlier items in a list can also be reduced when listening to synthetic speech (Luce *et al.*, 1983).

In summary, studies have shown that there is a wide range of performance between text-to-speech systems in terms of segmental intelligibility. Measured in terms of error rate, a system with a 3% error rate is twice as good as one with a 6% error rate, at least in terms of the average time interval between misperceptions in running text. Language is sufficiently redundant that these differences in segmental intelligibility often appear to be slight, but this is not the case when listening to unfamiliar names or difficult material. Furthermore, errors are usually the result of deviations of synthesizer parameters from values seen in natural speech. To the extent that error rate reflects a tendency for misspecification of parameters in general, it is also an indicator of how *unnatural* the speech is likely to sound.

TABLE XI. Performance of several text-to-speech systems with respect to listening comprehension (percent of questions about paragraph contents that were answered correctly), compared with visual presentation, after Pisoni and Hunnicutt, 1980).

Device	% correct	
Natural speech	68	
MITalk-79	70	(75% on second half of test)
Prose-2000	65	
Visual presentation	77	

D. Naturalness

Naturalness is a multi-dimensional subjective attribute that is not easy to quantify. Any of a large number of possible deficiencies can cause synthetic speech to sound unnatural to varying degrees. Fortunately, systems can be *compared* for relative subjective naturalness with a high degree of inter-subject and test–retest agreement (IEEE, 1969; Munson and Karlin, 1962). A standard procedure is to play pairs of test sentences synthesized by each system to be compared, and obtain judgments of preference (Logan and Pisoni, 1986). As long as the sentences being compared are the same, and the sentences are played without a long wait in between, valid data can be obtained. It is more difficult to compare systems that have been heard on different days or with different synthetic materials since extraneous factors can add an unpredictable amount of "noise" into listener preference judgment data (Nusbaum *et al.*, 1984).

Naturalness should not be confused with intelligibility. Some of the low bit rate linear-prediction systems sound like slightly distorted recordings of natural speech (which is what they are), and so are judged fairly natural, but they test out to have rather poor intelligibility scores (Nixon *et al.*, 1985). On the other hand, intelligibility and naturalness ratings of text-to-speech systems appear to be fairly highly correlated.

E. Suitability for a particular application

Text-to-speech devices are being introduced in a wide range of applications. A sampling of commercial uses appears in Table XII. Noncommercial applications are described in Sec. V. These devices are not good enough to fully replace a human, but they are likely to be well received by the general public if they are part of an application that offers a new service, or provides direct access to information stored on a computer, or permits easier or cheaper access to a present service because more telephone lines can be handled at a given cost. Both intelligibility and naturalness are considered important factors to the success of any application, but it is interesting to note that one large commercial concern is planning an application that will use DECtalk set up to speak in a monotone, purposely trying to indicate to the customer that he/she is talking to a smart computer rather than to a poor imitation of a human. What is important at this early stage in the exposure of the public to synthetic speech is to avoid applications that might lead to user frustration and generate negative attitudes toward all devices that "*talk like a computer.*" For example, intelligibility over the telephone

TABLE XII. Selected commercial applications for text-to-speech.

TEXT-TO-SPEECH BUSINESS APPLICATIONS
• Telephone information: e.g., 800 numbers for stock quotations, weather, ski conditions, sports scores, museum exhibits/schedules, talking Yellow Pages, ... (information that is changed frequently, and is available in computerized text form)
• Remote (on the road) access to computer mail
• Catalog ordering by phone, banking by phone (requires keypad or speech recognition for input)
• Data-base inquiry, especially for unsophisticated users: e.g., sales reps can determine status of purchase orders
• Generation of cassette recorded instructions for assembly plants, backplane wiring, telephone circuits, etc. (Flanagan *et al.*, 1972)
• Telephone access to computerized repair "experts" on, e.g., computers, telephone circuits.
• Coordination of large numbers of people on the road through a central computer information bank
• Warning and alarm systems concerning malfunctioning equipment
• Talking terminals and training devices (speech is often better than reading)
• Proofreading (catches kinds of typing errors that are often hard to detect visually)

of current text-to-speech systems may not be adequate for applications that involve many unfamiliar names (the telephone itself is somewhat marginal for this purpose, synthesis implies a further intelligibility reduction, and relatively poorer performance of these systems in converting names to phonemes adds additional potential confusion).

A current limitation for systems using the telephone is that the computer cannot listen as well as it can talk. Speech recognition technology is lagging behind synthesis capabilities. Presently, any user responses must be entered by telephone key pad commands, and not all telephones are push-button phones. Speaker-independent connected digit recognition systems are now being demonstrated in the laboratory with better than 98% string recognition accuracy (Bush and Kopek, 1986); perhaps this technology will become commercially available soon.

Eventually, text-to-speech systems will compete with products now used to produce canned messages from waveform-coding chips. Currently, such waveform encoding systems are thought to produce far more natural speech (due to natural timing, intonation, and voice quality obtained from a human utterance), even though measured intelligibility is significantly lower than for the better text-to-speech systems (Nixon *et al.*, 1985). If text-to-speech systems can be increased in naturalness even slightly, the advantages in terms of ease of message assembly can easily outweigh the cost advantage accruing to the waveform coder for many applications.

V. SPECIAL APPLICATIONS

Text-to-speech systems are beginning to be applied in many ways, including aids for the handicapped, medical aids, and teaching aids. This brief section is included in the hope of stimulating additional humanitarian applications for this new technology. It is an unfortunate fact of life in the United States that funds for transfer of this technology to the handicapped are scarce, and funds to actually purchase devices for individuals are virtually nonexistent. We depend on the success of text-to-speech systems in the commercial marketplace to lead to less costly portable low-power devices that, through the good will of industry, may be made available at special discounts for the handicapped.[20]

A. Talking aids for the vocally handicapped

The first kind of aid to be considered is a talking aid for the vocally handicapped. There are over 1.5 million nonspeaking persons in the USA, excluding the deaf, according to a survey made by the American Speech and Hearing Association (ASHA, 1981). Any person in this group who can point at some kind of a communication board or use a typewriter keyboard is a potential user of a communication aid that involves conversion of text to speech. A continually updated listing of communication aids for the nonvocal is maintained by the Trace Center of the University of Wisconsin (Vanderheiden, 1978, 1985); see also the quarterly publication *Communication Outlook* (Portnoy, 1979–present) and Bernstein (in press).

A potential advantage of DECtalk in this application is the possibility of fitting the voice characteristics to the user, particularly the advantage of giving women a femalelike voice and children a childlike voice. Prior to the availability of DECtalk, a 16-year old girl in Arizona who was injured in an automobile accident refused to use a talking aid because it made her sound masculine. On the other hand, some young cerebral palsy children seem to enjoy having a robotlike monotone voice speak for them when among their peers in a classroom setting.

Warrick *et al.* (1977) identify a number of capabilities that would facilitate wider use of talking aids: (1) natural distinguishable voices for each child in a classroom, (2) ability to express emphasis and attitude, (3) lighter weight and more portable configurations, (4) predictive type-ahead or other methods for speeding text specification. As pointed out by Bernstein (1986), natural voices are distinguished from one another by many types of cues that not only signal gender, but also approximate size, age, and regional accent. Current synthesis algorithms modify vocal tract size and laryngeal waveform to distinguish among a small set of speakers, but do not include capabilities to modify dialect, timing, intonation, allophonic selection, or phonetic realization. Users of talking aids can be frustrated by an inability to convey emotions such as urgency or friendliness by voice. Everything comes out in a sort of semantically neutral way, although some systems provide an ability to emphasize selected words.

The vocally handicapped present a wide range of motor difficulties that requires ingenious solutions to permit text creation. One method for speeding up text input is to use a predictive input system that always displays the most frequent English word for any typed word fragment, and the user can hit a special key to accept the prediction (Hunnicutt, 1985). Another alternative, similar in some ways to shorthand, is the Bliss symbol system (Carlson *et al.*, 1982b;

Hunnicutt, 1984). Each symbol stands for a common word. The authors found that Bliss symbols seem to be a good way to get nonvocal children started on language production, but a switch to normal orthography seems desirable later.

Vocal communication difficulties are fairly common when many prelingual deaf try to communicate with a hearing individual who is unfamiliar with the speech of the deaf. Bernstein *et al.* (1984) have designed a telephone communication system to overcome this problem. A deaf person can speak utilizing a Speech Plus text-to-speech board, and can "listen" by viewing the output of a large-vocabulary isolated-word speech recognition device. Preliminary data suggest that the performance of the recognition system in current use is marginal, but still good enough to be useful; future improvements could make such systems more attractive. An important issue is how to format the recognition alternatives, e.g., phonemes or a word hypothesis lattice (Huggins *et al.*, 1986).

B. Training aids

In one sense, a talking aid is by default a language training aid because it promotes practice and elicits direct feedback. This is one reason why it would be advantageous to get more talking aids to nonvocal children as early as possible in their school career. Experience suggests that this kind of device will also promote correct spelling and syntax (Carlson *et al.*, 1980). The inherent attraction of computer devices may mean that the approach could also be used with normal children for initial reading instruction.

A novel and quite successful application of text-to-speech is in the area of training dyslexic children to read. Dyslexia is a self-perpetuating difficulty because it is embarrassing to be helped by a teacher or friend, and it is nearly impossible to practice reading without help. Now several research groups have devised computer systems that permit unsupervised reading practice (Atkinson, 1972; Olson *et al.*, 1985). For example, the system being developed at the University of Colorado–Boulder (Olson *et al.*, 1985), uses a computer display screen, a mouse pointer, and a DECtalk text-to-speech system to read unfamiliar words or sound them out syllable by syllable.

Training aids need not be restricted to handicapped individuals (Sherwood, 1981). It is well known that speech has measurable advantages over reading and writing in many cognitive situations (Ochsman and Chapanis, 1974). For example, Suppes (1979, 1981) devised a computerized course for teaching algebra, and showed that providing some of the interactions via spoken responses resulted in better learning performance than visual presentation of all computer responses. Nakatani *et al.* (1986) provide spoken tutoring in the use of a text editor, noting that otherwise the student must constantly switch attention between the behavior of the editor and any tutorial information provided at the bottom of the screen. Tutorials of this sort can be made to have quite natural intonation and phrasing by proper annotation of the tutorial text (Hirshberg and Pierrehumbert, 1986).

C. Reading aids for the blind

Another application area is the development of reading aids for the blind. According to a survey by the National Center for Health Statistics (NCHS, 1977), approximately 1.4 million Americans are so severely visually impaired that they are unable to read ordinary news print, even with glasses. Machines that can scan printed material and produce speech would be of great help to this community. Ultimately, the goal is personal reading machines, although the current cost of the best performing machine, Kurzweil's, at a price of over $30 000, is far from this objective.

With the advent of computer typesetting, and the large text data bases available to computer users, it may not be necessary for a blind person to obtain a high-cost text reader. In some cases, connecting a text-to-speech system to a personal computer that is interfaced to a large network may serve many information gathering needs. One interesting pilot project in Sweden uses an FM overnight broadcast to load the day's newspaper into a blind person's personal computer, and has indexing programs to permit scanning of topics, using the Infovox SA-101 text-to-speech system (Carlson *et al.*, 1976; Carlson *et al.*, 1981).

Other efforts in this area have been concerned with timely production of talking books for the blind. Currently, most cassette recordings of books for the blind are produced by Recording for the Blind in Princeton, New Jersey. They use volunteer readers, and find that it takes up to 6 months after an order is placed to produce an audio copy of a typical textbook—primarily because of the problems inherent in coordinating volunteer efforts when so many hours of speech are to be produced. There are two possible ways to speed up the delivery of textbook orders. A text-to-speech system can work day and night to produce audio cassette tapes, or the text could be placed on a disk for a personal computer—in which case a blind person having a personal text-to-speech system could listen to the book and potentially be able to skim and scan over the book much more efficiently than is presently possible. An adjustable speaking rate, in conjunction with a computerized index or other method of content addressing could make reading almost as easy as the browsing we take for granted when we pick up a book (example 36 of the Appendix).

Another application is in the area of aids in the workplace. Of the 1.4 million visually impaired, many are elderly. However, 37 000 are children below 18, and 360 000 are between 18 and 64. Of these, about 106 000 are employed, according to the National Center for Health Statistics. Those blind individuals who work in an office environment could increase their productivity and become less dependent on a sighted co-worker if some sort of "talking text editor" computer system were available. Currently available systems are reviewed in *Aids and Appliances Review* (McGillivray, 1983).

D. Medical applications

Most medical applications are no different from other business applications, in that large health maintenance organizations employ centralized computer-based records on

patients that have to be accessed by phone when a doctor is not near a computer terminal. Attempts to use text-to-speech capabilities in novel ways have led to a computer system that tracks compliance in an experimental hypertension treatment program at Boston University Medical Center (Friedman, private communication). The computer calls each patient every day, and uses DECtalk to ask whether medication has been taken and whether any adverse side effects have occurred. The computer then calls a doctor if the patient's telephone keypad response indicates a problem.

Another potential application is an expert system for medical consultation between a doctor and a computerized data base. Those involved in artificial intelligence research have begun to amass large data bases on relations between symptoms and diseases. They hope ultimately to be able to reason logically, suggest additional tests, and deduce disease as well as the average practitioner—taking advantage of the superb memory capabilities of computers in order to consider rare clusters of symptoms that many doctors have not encountered in their practice. Text-to-speech telephone access could make such systems widely accessible and inexpensive.

VI. CONCLUSIONS

Text-to-speech conversion is a new technology with a rapidly changing set of capabilities and potential applications. The best of the current systems are quite intelligible, but suffer from a number of deficiencies that are often grouped under the catch-all term "lack of naturalness." In this article, we have identified many areas where rules and table values can be incrementally improved in the future to achieve more natural and more intelligible speech output from text-to-speech systems. As a consequence, these systems should become more acceptable to a wide range of users.

We have also identified several more basic problems that impede progress in certain areas of the text-to-speech conversion process (and also impact adversely on progress in other areas of speech science and technology). The first has to do with fitting spectral data obtained from female voices into the framework of current formant synthesizer models. For breathy vowels, the fit is not particularly good (recall Fig. 13), and it appears that some of the spectral deviations caused by tracheal coupling have perceptual importance.

It may be worthwhile to speculate on ways in which this problem might be resolved. Ideally, a new formant synthesizer model will be suggested that is slightly more complex, but still practical to implement. For example, an extra pole, or pole-zero pair might be made available to match extra spectral prominences that are observed. In this scenario, a way will be found to relate speech data from female voices to model parameters, so that a data collection effort will result in effective rules for controlling the new synthesizer model.

I suspect that the solution will not be that simple. If true, we may have to wait for speech science to provide better answers to some basic questions. The first point to note is that the acoustic theory of speech production, whether simplified or made complex by the introduction of better models of the larynx, trachea, and source–filter interactions, is not intended to be a model of the parameters directly controlled when we speak, nor of the parameters directly involved in the perceptual decoding of speech. The theory is a description of the acoustic behavior of a mechanical system. Therefore, efforts to relate observed spectral data from real female talkers to formant frequencies and other acoustic parameters of the theory have no *a priori* reason to succeed, and actually stand a good chance of failure, in part because there are too many model parameters compared with available spectral details (especially for talkers with high fundamental frequencies). Are we in a situation where it is possible to collect spectral data, yet be unable to relate it unambiguously to the underlying generation process, or to the processes of speech perception or articulation?

If this characterizes the present state of speech science, and I think it does, then the real bottleneck is the absence of a satisfactory perceptual theory to account for listeners' behavior in terms of observable spectral or waveform details. That we are far from such a theory is obvious, but how to go about attaining one is less clear. Attempts to mimic the steps believed to occur during the encoding stages of peripheral auditory processing are attractive as a first step, but it is unlikely that this encoding alone will be able to explain all of the fundamental perceptual skills that come naturally to humans, but not to speech recognition devices. Even the simplest of objectives, such as being able to categorize static critical-band spectra of vowels on the basis of a distance metric (Bladon and Lindblom, 1981), or to relate pairs of vowel spectra in terms of phonetic similarity (Klatt, 1982c), are well beyond our capabilities and understanding. Figure 34 shows pairs of critical-band spectra of vowels similar to/a/ that illustrate some of the difficulties encountered by a Euclidean metric. Spectral changes that affect peak locations are phonetically more important than other changes, even for low-pitched male voices synthesized to conform to the all-pole model of the vocal tract transfer function. But efforts to interpret critical-band spectra in terms of peak locations are thwarted in higher-pitched voices because individual harmonics are resolved, and breathy vowels introduce unexpected extra peaks. So long as we cannot always interpret spectral data from high-pitched voices in terms of formant parameters, or characterize the perceptual implications of spectral details, it is very likely that a synthetic female voice will remain an elusive goal, as may some aspects of the perceived naturalness of all male and female voices created by rules.

The second set of fundamental problems that we have identified arises when contemplating the creation of "natural" rule systems that manipulate articulatory structures. Where are the data that might facilitate creation of realistic models and model behavior? The acoustic consequences of any articulation depend on the cross-sectional area of the tube that is formed, and precision of specification is most important in locations of narrow constrictions. However, x-ray data, which are sparse, give only rough outlines in two dimensions, from which cross-sectional area must be inferred. And x-ray data do not characterize the masses and

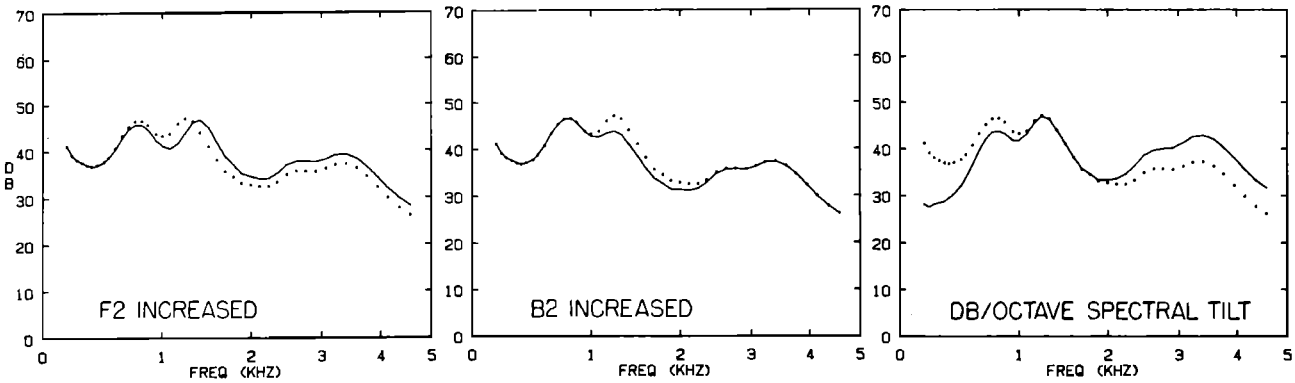

FIG. 34. Critical-band spectra of pairs of vowels that differ in terms of formant frequency location, formant bandwidth, or spectral tilt. Euclidean distance between solid and dotted curves does not reflect phonetic similarity between vowel pairs: An $F2$ increase creates a large change in judged phonetic difference, a B2 change is hard to hear at all, and a spectral tilt change is very audible, but does not affect judged phonetic similarity. Locations of energy concentrations seem to be of prime importance for phonetic categorization, but this hypothesis is difficult to maintain for high-pitched breathy vowels, see text.

degrees of freedom of the component articulators that make up the speech production system, nor the control constraints and strategies. There are a large number of researchers active in this area who make use of a wide range of devices to measure mechanical motions of individual articulators, and even EMG signals in individual muscles, but few if any of these scientists are pursuing a goal directly related to the assembly of synthesis rules for English syllables and sentences. Theorizing on the potential advantages of articulatory models in linguistics is another active area, but until such time as feature implementation rules become the central focus of effort (Goldstein and Browman, 1986), resulting in models detailed enough to reveal the immensity of the control problem, such theorizing is of marginal value to us. In summary, it seems that the study of basic processes of speech production and speech perception is crucial to progress, but is only in its infancy. Support of these activities will ultimately provide us with new insights and technical abilities. In the other direction, text-to-speech conversion is an excellent focus for sharpening the questions asked in basic research.

Let me quickly enumerate those areas where improvement in text-to-speech system performance is possible and reasonably straightforward. In all of the current systems, text analysis errors of many sorts are still possible. Deficiencies which have been identified in this article are summarized in Table XIII. The formatting routines may not be primed to deal with unusual letter or number strings. The word pronunciation routines have a certain probability of error in dealing with unfamiliar words, and this error rate tends to go up when dealing with foreign words and proper names. The syntax analysis routines may not be able to properly derive phrase structure for some sentences, or they may be unable to choose between two alternative pronunciations of an ambiguous orthographic word. These errors of text analysis are moderately frequent, occurring in as many as a third of the sentences of running text. Incremental improvements to formatting routines, augmentations to ever larger morpheme dictionaries (Coker, 1985), and additional parsing heuristics should lead to improved performance in this area. On the other hand, high performance syntactic analysis may turn out to require semantic knowledge, which would imply very large data structures and programs that may not be available for some time.

The problems remaining in the synthesis algorithms of text-to-speech systems are also listed in Table XIII. If one makes a spectrogram of a sentence produced by a text-to-speech system, and compares it with a sentence read by the person whose speech formed the basis for system development, it is easy to see ways in which the two acoustic patterns

TABLE XIII. Research issues for improving text-to-speech systems.

TEXT ANALYSIS
Text formatting
• programming what readers know about standard formats and abbreviations
Syntax/semantics
• syntactic analysis specifically for text-to-speech
• bootstrapping semantic information
Phoneme/stress prediction
• large-scale morphemic decomposition
• proper names
SYNTHESIS
Prosodics
• new systems of rules for f_0 control, improved duration rules
• mechanisms for getting variety into the rules
Phonology
• additional details concerning sentence-level phonetic recoding
Acoustic–phonetics
• segmental intelligibility
• detailed cues that may contribute to naturalness
Voice quality
• voicing source and tracheal coupling characterization for female voices
• source control as a function of time
APPLICATIONS
Technology transfer
• getting this technology into aids for the handicapped

differ. It is less easy to tell whether individual differences are perceptually important, but if one has some idea of discrimination limits, the perceptual salience of various speech cues, and the articulatory basis of acoustic discrepancies, then good guesses can be made as to the specific rules needed in the future. In this sense, all of the systems are amenable to incremental improvements so long as their designers have sufficient patience to follow this cookbook method of uncovering acoustic deficiencies.

Part of this process might even be automated. Holmes (1984) describes an effort to automatically time align a sentence with its synthetic imitation produced by rule, and then incrementally adjust formant frequency table values in the Holmes *et al.* (1964) rule program until natural and synthetic utterances are maximally similar. If the rules are correctly formulated and complete, such optimization procedures should result in improved imitations of other sentences as well. However, before such optimization efforts realize their full potential, many additional rules appear to be needed at the segmental level, e.g., to derive nuances of vowel quality change as a function of stress and phonetic environment. In the absence of a correct rule framework, automatic training will simply fail to converge, no matter how much data are supplied.

Text-to-speech programs and research may begin to have an influence on the way phonologists and phoneticians view phonetics and phonemic theory. These linguists have traditionally been reluctant to ascribe psychological reality to the phoneme, preferring to rely on distributional properties of observed sounds as a basis for theorizing (see Fry, 1974 for a good review). To the extent that speech generation programs begin to look like models of human behavior, their representations of language processes and units may become the cornerstones of new linguistic theories. If a synthesis-by-rule program can attract theoretical linguists to the problems inherent in specification of feature implementation rules, and thereby better couple their insights to the problem of allophonic variation, acoustic–phonetic detail, and timing of phonetic events, it is possible that real progress can be made in both engineering and linguistics. At the very least, it can be expected that these programs, in modifiable form, will become a part of the experimental facilities of modern phonetics laboratories, and will influence future generations of students in ways that are hard to predict.

In a similar vein, it is difficult to estimate the impact on the general public of computers that speak and listen. Talking machines may be just a passing fad, but the potential for new and powerful services is so great that this technology could have far reaching consequences, not only on the nature of normal information collection and transfer, but also on our attitudes toward the distinction between man and computer.

It is sometimes said that speech synthesis is not only *easier* than automatic speech recognition, but also that the field is so mature that the remaining problems are minor and scientifically uninteresting. I hope that this review has tended to dispel this view by pointing to specific areas where basic knowledge is lacking, and significant progress can still be made.

ACKNOWLEDGMENTS

Preparation of this review was supported in part by an NIH grant. I am very grateful to Ignatius Mattingly, John Holmes, Jared Bernstein, Osamu Fujimura, Stefanie Shattuck-Hufnagel, and David Pisoni for numerous suggestions based on an earlier draft.

APPENDIX: DEMONSTRATION

The enclosed 33 ⅓-rpm recording contains illustrations of some of the milestones in the development of systems for text-to-speech conversion. For convenience in locating and listening to examples as they are described in the text, it may be desirable to transfer the recording onto a cassette tape. The assistance of H. David Maxey, Michael Hecker, John Holmes, Patrick Nye, Joe Olive, and James Flanagan in assembling these materials is gratefully acknowledged. My thanks also go to Kenneth Stevens, who served as narrator.

The record has been inserted inside the back cover of this issue.

Part A: Development of speech synthesizers

The objective of early research on speech synthesis was to test whether the synthesizer design is capable of high-quality imitations of human voices.

1. The VODER of Homer Dudley, 1939.

Dudley of AT&T Bell Laboratories designed a speech synthesizer known as the "Voder" (Dudley *et al.*, 1939). It was demonstrated at the 1939 World's Fair in New York.

2. The Pattern Playback designed by Franklin Cooper, 1951.

The Haskins Laboratories Pattern Playback (Cooper *et al.*, 1951) was designed to permit converting back into sound the patterns observed on broadband sound spectrograms.

3. PAT, the "Parametric Artificial Talker" of Walter Lawrence, 1953.

Lawrence (1953) of the Signals Research and Development Establishment, Christchurch, England, designed the "PAT" ("Parametric Artificial Talker") parallel formant synthesizer. It was first demonstrated at a conference in London in 1952.

4. The "OVE" cascade formant synthesizer of Gunnar Fant, 1953.

Fant (1953) of the Royal Institute of Technology in Stockholm, Sweden designed a cascade formant synthesizer ("OVE I"). It was demonstrated at the same London conference in 1952.

5. Copying a natural sentence using Walter Lawrence's PAT formant synthesizer, 1962.

Tony Anthony and Walter Lawrence attempted to match a natural recording using an updated version of PAT (Anthony and Lawrence, 1962). Demonstrated at the 1962 Stockholm Speech Communication Conference. Compare with the OVE II version of the same utterance, next.

6. Copying the same sentence using the second generation of Gunnar Fant's OVE cascade formant synthesizer, 1962.

Gunnar Fant attempted to match a natural recording using OVE II (Fant and Martony, 1962). Demonstrated at the 1962 Stockholm Speech Communication Conference. Compare with the PAT version of the same utterance, above.

7. Comparison of synthesis and a natural sentence, using OVE II, by John Holmes, 1961.

Holmes (1961) of the Joint Speech Unit of the British Post Office used the OVE II synthesizer to generate a close copy of a natural sentence.

8. Comparison of synthesis and a natural sentence, John Holmes using his parallel formant synthesizer, 1973.

Holmes did essentially the same thing in 1973, using a more complex parallel formant synthesizer of his own design (Holmes, 1973). Demonstrated at the 1972 IEEE Conference on Speech Communication and Processing, Boston.

9. Attempt to scale the DECtalk male voice to make it sound female.

The DECtalk "Perfect Paul" male voice has been modified by scaling f_0 by a factor of 1.7 (ap = 204, pr = 170), by scaling all formant frequencies by a factor of 0.85 (hs = 85) and removing the fifth formant (f5 = 2500, b5 = 2048), by increasing the open quotient of the glottal waveform using the "richness" variable (ri = 0), and by decreasing the output level slightly to avoid overloads (lo = 81). These manipulations are not sufficient to turn Paul into a convincing female speaker.

10. Comparison of synthesis and a natural sentence, female voice, Dennis Klatt, 1986b.

A synthetic copy of a female speaker producing (1) a sentence and (2) an utterance in which each syllable of "Steve eats candy cane" is replaced by [ʔa] is compared with the original recording (Klatt, 1986b).

11. The DAVO articulatory synthesizer developed by George Rosen at M.I.T., 1958.

The DAVO ("Dynamic Analog of the VOcal tract") circuit designed by Rosen (1958) at M.I.T., augmented by a nasal tract designed by Hecker (1962), was controlled by a tape recording of control signals created by hand by Kenneth Stevens and Arthur House. The demonstration occurred at the fall meeting of the Acoustical Society of America in 1961.

12. Sentences produced by an articulatory model, James Flanagan and Kenzo Ishizaka, 1976.

Flanagan and Ishizaka (1976) of the AT&T Bell Telephone Laboratories used an articulatory synthesizer to generate two sentences, using control data derived from the Coker *et al.* (1973) text-to-speech system. A two-mass model of the vocal cords was employed, and turbulence noise was injected automatically whenever the Reynolds number became large at the larynx, or at a constricted section of the vocal tract.

13. Linear-prediction analysis and resynthesis of speech at a low-bit rate in the Texas Instruments Speak-'n-Spell toy, Richard Wiggins, 1980.

Wiggins (1980) designed a low-cost linear-prediction synthesis chip to take advantage of the ability of linear prediction to represent critical spectral and temporal aspects of speech waveforms efficiently.

14. Comparison of synthesis and a natural recording, automatic analysis–resynthesis using multipulse linear prediction, Bishnu Atal, 1982.

Atal of the AT&T Bell Laboratories demonstrated a new formulation of linear prediction, known as multipulse LPC (Atal and Remde, 1982) at the 1982 Paris ICASSP.

Part B: Segmental synthesis by rule

The first synthesis-by-rule programs concentrated on the development of rules for phonemic synthesis, and did not include rules for the automatic specification of phoneme durations and fundamental frequency. Since prosody was specified by hand to match a natural recording, these demonstrations sound significantly better than they would if all information had been derived by rule.

15. Creation of a sentence from rules in the head of Pierre Delattre, using the Haskins Pattern Playback, 1959.

A stylized spectrogram of the desired sentence was painted on a transparant plastic plate by Pierre Delattre, and then played by the Haskins Pattern Playback.

16. Output from the first computer-based phonemic synthesis-by-rule program, created by John Kelly and Louis Gerstman, 1961.

Kelly and Gerstman (1961, 1962) of the AT&T Bell Laboratories demonstrated the first phonemic synthesis-by-rule program in 1961 at a meeting of the Acoustical Society of America.

17. Elegent rule program for British English by John Holmes, Ignatius Mattingly, and John Shearme, 1964.

Holmes *et al.* (1964) of the Joint Speech Research Unit in England demonstrated an impressive phonemic synthesis-by-rule program for British English at the fall meeting of the Acoustical Society of America in Ann Arbor, 1963.

18. Formant synthesis using diphone concatenation, by Rex Dixon and David Maxey, 1968.

Dixon and Maxey (1968) of IBM at Research Triangle Park demonstrated a diphone concatenation method for construction of control parameter time functions for a formant synthesizer at the 1967 M.I.T. Conference on Speech Communication and Processing.

19. Rules to control a low-dimensionality articulatory model, by Cecil Coker, 1968.

Coker (1968) of AT&T Bell Laboratories created a method of generating speech from an articulatory model. The system was demonstrated at the 1967 M.I.T. Conference on Speech Communication and Processing.

Part C: Synthesis by rule of segments and sentence prosody

The next synthesis-by-rule programs include a complete set of rules for going from phonemes, stress marks, and some syntactic information to an output speech waveform.

20. First prosodic synthesis by rule, by Ignatius Mattingly, 1968.

The synthesis-by-rule program of Mattingly (1966; 1968) of the Haskins Laboratories was demonstrated to accompany his Ph.D. thesis.

21. Sentence-level phonology incorporated in rules by Dennis Klatt, 1976.

Klatt (1976b) of the M.I.T. Speech Communication Group created a phonological component to generate segmental durations and a fundamental frequency contour, as well as sentence-level allophonic variation, from a phonemic input augmented with stress and syntactic symbols.

22. Concatenation of linear-prediction diphones, by Joe Olive, 1977.

Olive (1977) of AT&T Bell Laboratories controlled a linear-prediction synthesizer from stored reflection coefficients for a set of diphones. The system was demonstrated at ICASSP-77. The recording is from about 1980, and includes prosodic rules provided by Liberman and Pierrehumbert.

23. Concatenation of linear-prediction demisyllables, by Cathrine Browman, 1980.

A synthesis-by-rule program with prosodic rules, called Lingua, was designed by Browman (1980) of AT&T Bell Laboratories, using the demisyllable inventory collected by Fujimura and Lovins (1978). Demonstrated at ICASSP-80.

Part D: Fully automatic text-to-speech conversion

24. The first full text-to-speech system, done in Japan by Noriko Umeda *et al.*, 1968.

The first demonstrated text-to-speech system for English was created by Umeda *et al.* (1968) of the Electrotechnical Laboratory in Japan, and was based on an articulatory model. It included a syntactic analysis module with sophisticated heuristics. Demonstrated at the 6th International Congress on Acoustics, in Tokyo in 1968.

25. The first Bell Laboratories text-to-speech system, by Cecil Coker, Noriko Umeda, and Cathrine Browman, 1973.

Coker *et al.* (1973) of AT&T Bell Laboratories demonstrated a text-to-speech program based on the Coker (1967) articulatory model. The system was demonstrated at the 1972 International Conference of Speech Communication and Processing in Boston.

26. The Haskins Laboratories text-to-speech system, 1973.

The Haskins Laboratories text-to-speech system (Cooper *et al.*, 1973) used the Mattingly (1968) phoneme-to-speech rules coupled with a large dictionary.

27. The Kurzweil reading machine for the blind, Raymond Kurzweil, 1976.

Kurzweil (1976) began selling a reading machine with an optical scanner in the late 1970s. The system was demonstrated on the CBS evening news.

28. The inexpensive Votrax Type-n-Talk system, by Richard Gagnon, 1978.

The Votrax low-cost Type-n-Talk text-to-speech system combines a single-chip synthesis-by-rule program and formant synthesizer (Gagnon, 1978) with a version of the Elovitz *et al.* (1976) letter-to-sound rules. It was demonstrated at the 1978 ICASSP Conference.

29. The Echo low-cost diphone concatenation system, about 1982.

The Echo low-cost text-to-speech system concatenates linear-prediction diphones using the Texas Instrument's TMS-5220 linear prediction synthesizer chip.

30. The M.I.T. MITalk system, by Jonathan Allen, Sheri Hunnicutt, and Dennis Klatt, 1979.

The MITalk-79 laboratory text-to-speech system, developed at the Massachusetts Institute of Technology by Allen *et al.* (1979, 1987) and many others. The system was demonstrated in its final form at the 1979 meeting of the Acoustical Society of America in Boston.

31. The multi-language Infovox system, by Rolf Carlson, Bjorn Granström, and Sheri Hunnicutt, 1982.

The Infovox commercial text-to-speech system (Magnesson *et al.*, 1984) is an implementation of the Carlson *et al.* (1982a) multilanguage system that was developed at the Royal Institute of Technology in Stockholm by Rolf Carlson *et al.* Versions of the system were demonstrated in 1976 and 1982 at ICASSP conferences.

32. The Speech Plus Inc. "Prose-2000" commercial system, 1982.

The Prose-2000 commercial text-to-speech system was first developed in conjunction with a reading machine for the blind project at Telesensory Systems by James Bliss and his associates (Goldhor and Lund, 1983; Groner *et al.*, 1982). The recording is of Version 3.0 of the software.

33. The Klattalk system, by Dennis Klatt of M.I.T. which formed the basis for Digital Equipment Corporation's DECtalk commercial system, 1983.

The Klattalk (1982a) laboratory text-to-speech system software was licensed to Digital Equipment Corporation as a basis for the commercial DECtalk text-to-speech system announced in 1983. The recording is of Version 3.0 of the DECtalk software.

34. The AT&T Bell Laboratories text-to-speech system, 1985.

A new AT&T Bell Laboratories laboratory text-to-speech system (Olive and Liberman, 1985) uses the Olive (1977) diphone synthesis strategy in combination with a large morpheme dictionary (Coker, 1985) and letter-to-sound rules (Church, 1985). The laboratory system was demonstrated at a 1985 meeting of the Acoustical Society of America.

35. Several of the DECtalk voices.

Examples of some of the voices provided by the *DECtalk* text-to-speech system: (1) Beautiful Betty, (2) Huge Harry, (3) Kit the Kid, (4) Whispering Wendy.

36. DECtalk speaking at about 300 words/minute.

Example of using the DECtalk speaking rate command to skim material at a rapid rate. The nominal speaking rate has been set to 350 words/min, [:ra 350], although this 51-word passage took 11 s to speak, indicating an effective rate slightly under 300 words/min.

[1]For an application requiring a limited set of sentences with known struc-

ture, such as telephone numbers, some success has been achieved in concatenating "*vocoded*" whole words. This is because it is possible to smooth vocoder parameters at word boundaries, modify durations, and impose a sentence fundamental frequency contour on the word string (Rabiner *et al.*, 1971; Olive and Nakatani, 1974). Also, Cooper *et al.*, 1984) describe an early plan to concatenate recorded words in a reading machine for the blind application, where the motivation of the listeners might overcome weaknesses of the presentation, but the approach was subsequently abandoned in favor of synthesis by rule.

[2]Postvocalic devoicing and flapping are actually late rules, occurring after vowel durations are computed. The proper ordering of rules is an important issue in the design of text-to-speech systems.

[3]Looking at the same data, we might not agree with their intuitions.

[4]The parameter k was assumed to be 0.5 by Delattre *et al.* (1955) and by Holmes *et al.* (1964).

[5]Actually, Peterson *et al.* (1958) proposed the term "diad" as a set of diphones all having essentially the same articulatory trajectory from the middle of one segment to the middle of the next, but differing in prosodic values such as duration and fundamental frequency contour. Hank Truby was the first to use the term "diphone" by separating out prosody as independent variables in synthesis, and calling the remaining phonetic transition (as represented by synthesizer control data) a "diphone." As the term diphone has spread in usage, some authors allow it to refer to larger synthesis units such as consonant clusters when needed to maximize synthesis fidelity (Dixon and Maxey, 1968), but we will restrict the term here to mean a transition between adjacent phonetic segments.

[6]For example, English vowels can be divided into tense inherently long vowels and lax short vowels (House, 1961).

[7]The number of distinguishable stress levels at the lexical and phrasal levels continues to be an area of linguistic dispute; see Vanderslice and Ladefoged (1971) and Coker *et al.* (1973) for extremal positions.

[8]In phonological theory, there is usually a distinction made between a rule that changes a feature or segment discretely, and a feature implementation algorithm that is subject to low-level physiological constraints, contextual influences, and graded behavior. Thus a parameter adjustment rule needed in speech synthesis probably should correspond to the feature implementation level of description (e.g., voice onset time is slightly longer for high versus low vowels even though glottal timing commands might be the same in two situations), whereas allophone selection rules should correspond to actual rule-governed changes to motor commands, as reflected by a change to some segmental feature.

[9]Not all phonological simplifications preserve boundary information; for example [h] deletion and flapping result in an inability to distinguish between "but her" and "butter."

[10]If errors were independent, words correct would be approximately equal to phonemes correct to the sixth or seventh power, times the probability of getting the stress correct.

[11]It is perhaps unfair to evaluate this system against a random sample of words because it was intended to be used in the context of a large morpheme dictionary, and therefore would be activated only for rare words—words that may be more regular in their pronunciation.

[12]Use of the solid curve is equivalent to assuming that another million-word text sample will contain exactly the same 50 000 words, whereas it is likely that a different set of rare words will be found in the new text.

[13]It is surprising how outdated this corpus has become if the goal is to obtain a lexicon representative of modern textual material; Allen and Finkel removed more than 15% of the items as outmoded or too parochial when they were collecting morphemes by hand. We would all benefit from a modern replication of the Kučera and Francis task, especially now that it is practical to examine much larger data bases than only a million words.

[14]In theory, every time a new rule was added to the morph decomposition process, it was necessary to go back and check the entire lexicon for accidental incorrect decompositions.

[15]An even less sensitive test is the diagnostic rhyme test (Voiers, 1983) which involves a single pair of alternative responses for each familiar CVC word.

[16]Most of these "errors" can be attributed to problems with phonemic symbolization; phonetically trained listeners typically perform at better than 99% correct on the same task (Rabiner, 1969).

[17]Multipulse linear prediction was designed to make possible the detailed modeling of the voicing source waveform, but in fact it is simply a method of introducing zeros into the representation of any speech sound. It appears that multipulse has little advantage for voiced segments in text-to-speech systems because the rule system imposes an f_0 contour different from that observed in the original natural speech recording. However, multipulse may be able to better approximate, e.g., the coherent release of plosive bursts (Maeda, 1987).

[18]Pisoni and Koen (1981) obtained similar results, although the difference between natural and synthetic speech was greater, perhaps because the MITalk system that they used is not quite as intelligible.

[19]Carlson and Granström (1976) had noted the same kind of listener adaptation without feedback in an earlier experimental evaluation. With feedback, listeners can improve considerably in performance on intelligibility tests, even with poor quality synthetic speech (Schwab *et al.*, 1986).

[20]For example, Xerox Corp. has retrofitted a number of Kurzweil Reading Machines for the blind that are located in public libraries with the more intelligible Prose-2000 text-to-speech board. Digital Equipment Corporation has offered a special price for DECtalk units sold to handicapped individuals and manufacturers of handicapped devices, resulting in a more than one million dollar price reduction on units sold to this population.

Aho, A., and Ullman, J. (**1972**). *The Theory of Parsing, Translation and Computing* (Prentice-Hall, New York).

Ainsworth, W. A. (**1973**). "A System for Converting English Text into Speech," IEEE Trans. Audio Electroacoust. **AU-21**, 288–290.

Allen, D. R., and Strong, W. J. (**1985**). "A Model for the Synthesis of Natural Sounding Vowels," J. Acoust. Soc. Am. **78**, 58–69.

Allen, J. (**1976**). "Synthesis of Speech from Unrestricted Text," Proc. IEEE **64**, 422–433.

Allen, J., Hunnicutt, S., and Klatt, D. H. (**1987**). *From Text to Speech: The MITalk System* (Cambridge U.P., Cambridge, UK).

Allen, J., Hunnicutt, S., Carlson, R., and Granström, B. (**1979**). "MITalk-79: The MIT Text-to-Speech System," J. Acoust. Soc. Am. Suppl. 1 **65**, S130.

Ananthapadmanabha, T. V. (**1984**). "Acoustic Analysis of Voice Source Dynamics," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, **QPSR 2-3**, 1–24.

Anderson, M., Pierrehumbert, J., and Liberman, M. (**1984**). "Synthesis by Rule of English Intonation Patterns," Proc. Int. Conf. Acoust., Speech Signal Process. **ICASSP-84**, 2.9.1–2.9.4.

Anthony, J., and Lawrence, W. (**1962**). "A Resonance Analogue Speech Synthesizer," Proc. 4th Int. Cong. Acoust., Copenhagen, Denmark.

Armstrong, L. E., and Ward, I. C. (**1931**). *A Handbook of English Intonation* (Cambridge U.P., Cambridge, England), 2nd ed.

ASHA (**1981**). *Position Paper of the Ad Hoc Committee on Communication Processes for Nonspeaking Persons*, American Speech and Hearing Association, Rockville, MD.

Atal, B. S., and Hanauer, S. L. (**1971**). "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. **50**, 637–655.

Atal, B. S., and Remde, J. R. (**1982**). "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," Proc. Int. Conf. Acoust., Speech Signal Process. **ICASSP-82**, 614–617.

Atal, B. S., and Schroeder, M. R. (**1975**). "Recent Advances in Predictive Coding: Applications to Speech Synthesis," in *Speech Communication*, edited by G. Fant (Almqvist and Wiksell, Uppsala, Sweden), Vol. I, pp. 27–31.

Atkinson, R. C. (**1972**). "Teaching Children to Read Using a Computer," Am. Psychol. **27**, 169–178.

Baer, T. (**1981**). "Observation of Vocal Fold Vibration: Measurement of Excised Larynges," in *Vocal Fold Physiology*, edited by K. N. Stevens and M. Hirano (University of Tokyo Press, Tokyo, Japan), pp. 119–136.

Barnwell, T. P. (**1971**). "An Algorithm for Segment Duration in a Reading Machine Context," Research Laboratory of Electronics, Tech. Report 479, M.I.T., Cambridge, MA.

Bassak, G. (**1980**). "Phoneme-Based Speech Chip Needs Less Memory," Electronics **53**, 43–44.

Bell, T. (**1983**). "Talk to Me," Personal Computing **7**, 120–206 (September 1983).

Bernstein, J. (**1986**). "Voice Identity and Attitude," *Proceedings Speech Tech '86* (Media Dimensions, New York), pp. 213–215.

Bernstein, J., and Baldwin, G. (1985). "Spontaneous versus Prepared Speech," J. Acoust. Soc. Am. Suppl. 1 78, S37.
Bernstein, J., Becker, R., Bell, D., Murveit, H., Poza, P., and Stevens, G. (1984). "Telephone Communication between Deaf and Hearing Persons," Proc. Int. Conf. Acoust., Speech Signal Process. ICASSP-84, 26.7.1–26.7.4.
Bernstein, J., and Nessly, L. (1981). "Performance Comparison of Component Algorithms for the Phonemicization of Orthography," Proc. 19th Annu. Assoc. Computational Linguistics, Stanford Univ., pp. 19–22.
Bernstein, J., and Pisoni, D. B. (1980). "Unlimited Text-to-Speech System: Description and Evaluation of a Microprocessor-Based Device," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-80, 576–579.
Bernstein, L. (Ed.) (in press). *The Vocally Impaired* (Academic, New York).
Bickley, C. (1982). "Acoustic Analysis and Perception of Breathy Vowels," Speech Communication Group Working Papers I, MIT, Cambridge, MA, pp. 71–82.
Bilger, R., Nuetzel, J., Rabinowitz, W., and Rzeczkowski, C. (1984). "Standardization of a Test of Speech Perception in Noise," J. Speech Hearing Res. 27, 32–48.
Bladon, A., and Lindblom, B. (1981). "Modeling the Judgement of Vowel Quality Differences," J. Acoust. Soc. Am. 69, 1414–1422.
Bloomfield, L. (1933). *Language* (Henry Holt, New York).
Bolinger, D. (1951). "Intonation: Levels versus Configurations," Word 7, 199–210.
Bolinger, D. (1972). "Accent Is Predictable if You Are a Mind-Reader," Language 48, 633–644.
Broad, D. J. (1979). "The New Theories of Vocal Fold Vibration," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. Lass (Academic, New York), Vol. 2, pp. 203–256.
Broad, D. J., and Fertig, R. H. (1970). "Formant Frequency Trajectories in Selected CVC Syllable Nuclei," J. Acoust. Soc. Am. 47, 1572–1582.
Browman, C. P. (1980). "Rules for Demisyllable Synthesis using Lingua, a Language Interpreter," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-80, 561–564.
Bruckert, E., Minow, M., and Tetschner, W. (1983). "Three-Tiered Software and VLSI Aid Developmental System to Read Text Aloud," Electronics (21 April 1983).
Bush, M. A., and Kopek, G. E. (1986). "Network-Based Connected Digit Recognition Using Explicit Acoustic-Phonetic Modeling," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-86, 1097–1100.
Carlisle, J. F. (1985). "The Relationship between Knowledge of Derivational Morphology and Spelling Ability in Fourth, Sixth and Eighth Graders," Status Report on Speech Research SR-82/83, Haskins Laboratories, New Haven, CT, pp. 151–174.
Carlson, R., Galyas, K., Granström, B., Pettersson, M., and Zachrisson, G. (1980). "Speech Synthesis for the Non-Vocal in Training and Communication," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR-1, 13–27.
Carlson, R., Galyas, K., Granström, B., Hunnicutt, S., Larsson, B., and Neovius, L. (1981). "A Multi-Language Portable Text-to-Speech System for the Disabled," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR 2/3, 8–16.
Carlson, R., and Granström, B. (1975). "A Phonetically Oriented Programming Language for Rule Description of Speech," in *Speech Communication*, edited by G. Fant (Almqvist and Wiksell, Uppsala, Sweden), Vol. 2, pp. 245–253.
Carlson, R., and Granström, B. (1976). "A Text-to-Speech System Based Entirely on Rules," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-76, 686–688.
Carlson, R., Granström, B., and Hunnicutt, S. (1982a). "A Multi-Language Text-to-Speech Module," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-82, 1604–1607.
Carlson, R., Granström, B., and Hunnicutt, S. (1982b). "Bliss Communication with Speech or Text Output," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-82, 747–750.
Carlson, R., Granström, B., and Klatt, D. H. (1979). "Some Notes on the Perception of Temporal Patterns in Speech," 9th International Congress of Phonetic Sciences, Copenhagen, Denmark.
Carlson, R., Granström, B., and Larssen, K. (1976). "Evaluation of a Text-to-Speech System as a Reading Machine for the Blind," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR 2-3, 9–13.
Carlson, R., Granström, B., and Pauli, S. (1972). "Perceptual Evaluation of Segmental Cues," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR 1/1972, 18–24.
Carrell, T. D. (1984). "Contributions of Fundamental Frequency, Formant Spacing, and Glottal Waveform to Talker Identification," Speech Research Laboratory Progress Report 5, Indiana Univ., Bloomington, IN.
Catford, J. G. (1977). *Fundamental Problems in Phonetics* (Indiana U.P., Bloomington, IN).
Chial, M. (1985). "Tutorial on Speech Synthesis," American Speech-Language-Hearing Foundation 1985 Computer Conference, New Orleans, LA.
Chiba, T., and Kajiyama, M. (1941). *The Vowel: Its Nature and Structure* (Tokyo–Kaiseikan, Tokyo, Japan).
Chomsky, N., and Halle, M. (1968). *Sound Pattern of English* (Harper and Row, New York).
Church, K. (1983). "Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints," unpublished Ph.D. thesis, M.I.T., Cambridge, MA.
Church, K. W. (1985). "Stress Assignment in Letter-to-Sound Rules for Speech Synthesis," Proc. 23rd Meeting Assoc. Comp. Ling. 246–253.
Clements, G. N. (1985). "The Geometry of Phonological Features," Phonology 2, 223–252.
Cochran, P.S. (1986). "Comparing the Intelligibility of Speech Synthesizers," Proc. Am. Voice Input/Output Soc. AVIOS-86, 464–474.
Cohen, A., and 't Hart, J. (1967). "On the Anatomy of Intonation," Lingua 19, 177–192.
Cohen, P., and Mercer, R. (1974). "The Phonological Component of an Automatic Speech Recognition System," in *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*, edited by R. Reddy (IEEE, New York), pp. 275–320.
Coker, C. H. (1968). "Speech Synthesis with a Parametric Articulatory Model," Speech Symposium, paper A-4, Kyoto, reprinted in *Speech Synthesis*, edited by J. L. Flanagan and L. R. Rabiner (Dowden, Hutchinson and Ross, Stroudsburg, PA), pp. 135–139.
Coker, C. H. (1976). "A Model of Articulatory Dynamics and Control," Proc. IEEE 64, 452–459.
Coker, C. H. (1985). "A Dictionary-Intensive Letter-to-Sound Program," J. Acoust. Soc. Am. Suppl. 1 78, S7.
Coker, C. H., and Umeda, N. (1975). "The Importance of Spectral Detail in Initial-Final Contrasts of Voiced Stops," J. Phonetics 3, 63–68.
Coker, C. H., Umeda, N., and Browman, C. P. (1973). "Automatic Synthesis from Ordinary English Text," IEEE Trans. Audio Electroacoust. AU-21, 293–297.
Conroy, D., and Vitale, T. (1986). Personal communication.
Conroy, D., Vitale, T., and Klatt, D. H. (1986). "DECtalk DTC03 Text-to-Speech System Owner's Manual," EK-DTC03-OM-001, Educational Services of Digital Equipment Corporation, P.O. Box CS2008, Nashua, NH 03061.
Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (1952). "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am. 24, 597–606.
Cooper, F. S., Gaitenby, J. H., Mattingly, I. G., Nye, P. W., and Sholes, G. N. (1973). "Audible Outputs of Reading Machines for the Blind," Status Report on Speech Research SR-35/36, Haskins Laboratories, New Haven, CT, pp. 117–120.
Cooper, F. S., Gaitenby, J. H., and Nye, P. W. (1984). "Evolution of Reading Machines for the Blind: Haskins Laboratories' Research as a Case History," J. Rehab. Res. Develop. 21, 51–87.
Cooper, F. S., Liberman, A. M., and Borst, J. M. (1951). "The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech," Proc. Natl. Acad. Sci. (US) 37, 318–325.
Cooper, W. A., Paccia, J. M., and Lapointe, S. G. (1978). "Hierarchical Coding in Speech Timing," Cognitive Psychol. 10, 154–177.
Cooper, W. E., and Sorenson, J. (1981). *Fundamental Frequency in Sentence Production* (Springer, New York).
Dedina, M. J., and Nusbaum, H. C. (1986). "Pronounce: A Program for Pronunciation by Analogy," Speech Research Laboratory Progress Report 12, Indiana University, Bloomington, IN, pp. 335–348.
Davidsen-Nielsen, N. (1974). "Syllabification of English Words with Medial sp, st, sk," J. Phonetics 2, 15–45.
Delattre, P., Liberman, A. M., and Cooper, F. S. (1955). "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am. 27, 769–774.
Dixon, N. R., and Maxey, H. D. (1968). "Terminal Analog Synthesis of

Continuous Speech Using the Diphone Method of Segment Assembly," IEEE Trans. Audio Electroacoust. **AU-16**, 40–50.

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (**1977**). "Stop Consonant Recognition: Release Bursts and Formant Transitions as Functionally Equivalent Context-Dependent Cues," Percept. Psychophys. **22**, 109–122.

Dubno, J. R., and Levitt, H. (**1981**). "Predicting Consonant Confusions from Acoustic Analysis," J. Acoust. Soc. Am. **69**, 249–261.

Dudley, H. (**1939**). "The Vocoder," Bell Labs Rec. **17**, 122–126.

Dudley, H., Riesz, R. R., and Watkins, S. A. (**1939**). "A Synthetic Speaker," J. Franklin Inst. **227**, 739–764.

Dudley, H., and Tarnoczy, T. H. (**1950**). "The Speaking Machine of Wolfgang Kempelen," J. Acoust. Soc. Am. **22**, 151–166.

Dunn, H. K. (**1950**). "The Calculation of Vowel Resonances, and Electrical Vocal Tract," J. Acoust. Soc. Am. **22**, 740–753.

Egan, J. P. (**1948**). "Articulation Testing Methods," Laryngoscope **58**, 955–991.

Elovitz, H., Johnson, R., McHugh, A., and Shore, J. (**1976**). "Letter-to-Sound Rules for Automatic Translation of English Text to Phonetics," IEEE Trans. Acoust. Speech Signal Process. **ASSP-24**, 446–459.

Erber, N. P. (**1979**). "An Approach to Evaluating Auditory Speech Perception Ability," Volta Rev. **80**, 344–350.

Estes, S. E., Kerby, H. R., Maxey, H. D., and Walker, R. M. (**1964**). "Speech Synthesis from Stored Data," IBM J. Res. Develop. **8**, 2–12.

Fairbanks, G. (**1958**). "Test of Phonemic Differentiation: the Rhyme Test," J. Acoust. Soc. Am. **30**, 596–600.

Fant, G. (**1953**). "Speech Communication Research," Ing. Vetenskaps Akad. Stockholm, Sweden **24**, 331–337.

Fant, G. (**1956**). "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," in *For Roman Jakobson*, edited by M. Halle (Mouton, 's-Gravenhage, The Netherlands), pp. 109–120.

Fant, G. (**1959**). "Acoustic Analysis and Synthesis of Speech with Applications to Swedish," Ericsson Technics **15**, 3–108.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, 's-Gravenhage, The Netherlands).

Fant, G. (**1973**). "Stops in CV Syllables," in *Speech Sounds and Features*, edited by G. Fant (MIT Press, Cambridge, MA), pp. 110–139.

Fant, G. (**1975**). "Non-Uniform Vowel Normalization," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm **QPSR 2-3**, 1–19.

Fant, G. (**1979**). "Glottal Source and Excitation Analysis," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden **QPSR 1**, 85–107.

Fant, G. (**1982**). "Preliminaries to Analysis of the Human Voice Source," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden **QPSR 4**, 1–25.

Fant, G., and Ananthapadmanabha, T.V. (**1982**). "Truncation and Superposition," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden **QPSR 2-3**, 1–17.

Fant, G., Lin, Q. G., and Gobl, C. (**1985**). "Notes on Glottal Flow Interaction," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm **QPSR 2-3**, 21–45.

Fant, G., and Martony, J. (**1962**). "Speech Synthesis," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden **QPSR 2**, 18–24.

Fant, G., Stalhammer, U., and Karlsson, I. (**1974**). "Swedish Vowels in Speech Materials of Various Complexity," in *Speech Communication*, edited by G. Fant (Almqvist and Wiksell, Uppsala, Sweden), Vol. 2, pp. 139–147.

Fischer, F. W., Shankweiler, D., and Liberman, I. Y. (**1985**). "Spelling Proficiency and Sensitivity to Word Structure," J. Memory Lang. **24**, 423–441.

Fischer-Jorgensen, E. (**1954**). "Acoustic Analysis of Stop Consonants," Miscel. Phonetica **2**, 42–59.

Flanagan, J. L. (**1957**). "Note on the Design of Terminal Analog Speech Synthesizers," J. Acoust. Soc. Am. **29**, 306–310.

Flanagan, J. L. (**1958**). "Some Properties of the Glottal Sound Source," J. Speech Hearing Res. **1**, 99–116.

Flanagan, J. L. (**1972**). *Speech Analysis Synthesis and Perception* (Springer, New York).

Flanagan, J. L. (**1976**). "Computers that Talk and Listen: Man-Machine Communication by Voice," Proc. IEEE **64**, 405–415.

Flanagan, J. L. (**1981**). "Synthesis and Recognition of Speech: How Computers Talk," Bell Labs. Rec. **59** (April), 123–130.

Flanagan, J. L., and Ishizaka, K. (**1976**). "Automatic Generation of Voiceless Excitation to a Vocal Cord–Vocal Tract Speech Synthesizer," IEEE Trans. Acoust. Speech Signal Process. **ASSP-24**, 163–170.

Flanagan, J. L., and Ishizaka, K. (**1978**). "Computer Model to Characterize the Air Volume Displaced by the Vibrating Vocal Cords," J. Acoust. Soc. Am. **63**, 1558–1563.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (**1975**). "Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract," Bell Syst. Tech. J. **54**, 485–506.

Flanagan, J. L., and Landgraf, L. (**1968**). "Self-Oscillating Source for Vocal-Tract Synthesizers," IEEE Trans. Audio Electroacoust. **AU-16**, 57–64.

Flanagan, J. L., and Rabiner, L. R. (**1973**). *Speech Synthesis* (Dowden, Hutchinson and Ross, Stroudsburg, PA).

Flanagan, J. L., Rabiner, L. R., Schafer, R. W., and Denman, J. D. (**1972**). "Wiring Telephone Apparatus from Computer Generated Speech," Bell System Tech. J. **51**, 391–397.

Francis, W. N., and Kucera, H. (**1982**). *Frequency Analysis of English Usage* (Houghton Mifflin, Boston, MA).

Friedman, R. H. (**1986**). Private communication.

Fromkin, V. (**1971**). "The Non-anomalous Nature of Anomalous Utterances," Language **47**, 27–52.

Fry, D. B. (**1958**). "Experiments in the Perception of Stress," Lang. Speech **1**, 126–152.

Fry, D. B. (**1965**). "The Dependence of Stress Judgements on Vowel Formant Structure," in *Proceedings of the 6th International Congress of Phonetic Sciences*, edited by E. Zwirner and W. Bethge (Karger, Basel), pp. 306–311.

Fry, D. B. (**1974**). "Phonetics in the Twentieth Century," in *Current Trends in Linguistics*, edited by T. S. Sebeok (Mouton, 's-Gravenhague, The Netherlands), Vol. 12, pp. 2201–2239.

Fujimura, O. (**1960**). "Spectra of Nasalized Vowels," Research Laboratory of Electronics QPR 62, MIT, Cambridge, MA, pp. 214–218.

Fujimura, O. (**1962**). "Analysis of Nasal Consonants," J. Acoust. Soc. Am. **34**, 1865–1875.

Fujimura, O. (**1968**). "An Approximation to Voice Aperiodicity," IEEE Trans. Audio Electroacoust. **AU-16**, 68–72.

Fujimura, O., and Kakita, Y. (**1979**). "Remarks on Quantitative Description of the Lingual Articulation," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, New York), pp. 17–24.

Fujimura, O., and Lovins, J. (**1978**). "Syllables as Concatenative Phonetic Elements," in *Syllables and Segments*, edited by A. Bell and J. B. Hooper (North-Holland, New York), pp. 107–120.

Fujisaki, H., and Nagashima, S. (**1969**). "Synthesis of Pitch Contours of Connected Speech," Annual Report Engineering Res. Inst. Univ. Tokyo, Tokyo, Japan **28**, 53–60.

Fujisaki, H. (**1986**). "Proposal and Evaluation of Models for the Glottal Source Waveform," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-86**, 1609–1612.

Gagnon, R. T. (**1978**). "Votrax Real Time Hardware for Phoneme Synthesis of Speech," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-78**, 175–178.

Gaitenby, J. (**1965**). "The Elastic Word," Status Report on Speech Research SR-2, Haskins Laboratories, New Haven, CT, pp. 1–12.

Gaitenby, J. H., Sholes, G. N., and Kuhn, G. M. (**1972**). "Word and Phrase Stress by Rule for a Reading Machine," Status Report on Speech Research SR-29/30, Haskins Laboratories, New Haven, CT, pp. 105–110.

Gerstman, L. (**1957**). "Perceptual Dimensions for the Fricative Noise Portions of Certain Speech Sounds," unpublished Ph.D. dissertation, New York University, New York.

Glushko, R. J. (**1981**). "Principles for Pronouncing Print: The Psychology of Phonography," in *Interactive Processes in Reading*, edited by A. M. Lesgold and C. A. Perfetti (Erlbaum, Hillsdale, NJ).

Gold, B., and Rabiner, L. R. (**1968**). "Analysis of Digital and Analog Formant Synthesizers," IEEE Trans. Audio Electroacoust. **AU-16**, 81–94.

Goldhor, R. S., and Lund, R. T. (**1983**). "University-to-Industry Advanced Technology Transfer: A Case Study," Res. Policy **12**, 121–152.

Goldman-Eisler, F. (**1968**). *Psycholinguistics: Experiments in Spontaneous Speech* (Academic, New York).

Goldstein, L., and Browman, C. P. (**1986**). "Representation of Voicing Contrasts Using Articulatory Gestures," J. Phonetics **14**, 339–342.

Groner, G. F., Bernstein, J., Ingber, E., Pearlman, J., and Toal, T. (**1982**). "A Real-Time Text-to-Speech Converter," Speech Technol. **1**, 73–76.

Haggard, M. (1973). "Abbreviation of Consonants in English Pre- and Post-Vocalic Clusters," J. Phonetics 1, 9–24.
Halle, M. (1985). "Speculations about the Representation of Words in Memory," in *Phonetic Linguistics*, edited by V. Fromkin (Academic, New York), pp. 101–114.
Halle, M., Hughes, G. W., and Radley, J. P. A. (1957). "Acoustic Properties of Stop Consonants," J. Acoust. Soc. Am. 20, 107–116.
Halle, M., and Keyser, S. J. (1971). *English Stress: Its Form, its Growth and its Role in Verse* (Harper and Row, New York).
Halle, M., and Vergnaud, J. R. (1980). "Three-Dimensional Phonology," J. Ling. Res. 1, 83–105.
Halliday, M. A. K. (1970). "Functional Diversity in Language as seen from a Consideration of Modality and Mood in English," Found. Lang. 6, 322–361.
Harris, C. M. (1953). "A Study of the Building Blocks of Speech," J. Acoust. Soc. Am. 25, 962–969.
Harris, K. (1958). "Cues for Discrimination of American English Fricatives in Spoken Syllables," Lang. Speech 1, 1–17.
Harris, M. O., Umeda, N., and Bourne, J. (1981). "Boundary Perception in Fluent Speech," J. Phonetics 9, 1–18.
't Hart, J., and Cohen, A. (1973). "Intonation by Rule: A Perceptual Quest," J. Phonetics 1, 309–327.
Hecker, M. H. L. (1962). "Studies of Nasal Consonants with an Articulatory Speech Synthesizer," J. Acoust. Soc. Am. 34, 179–188.
Heffner, R. S. (1969). *General Phonetics* (Univ. Wisconsin, Madison, WI).
Heinz, J. M., and Stevens, K. N. (1961). "On the Properties of Voiceless Fricative Consonants," J. Acoust. Soc. Am. 33, 589–596.
Henderson, L. (1982). *Orthography and Word Recognition in Reading* (Academic, New York).
Henke, W. L. (1967). "Preliminaries to Speech Synthesis Based upon an Articulatory Model," Proc. IEEE Conf. Speech Commun. Process. 170–182.
Hertz, S. (1982). "From Text to Speech with SRS," J. Acoust. Soc. Am. 72, 1155–1170.
Hertz, S., Kadin, J., and Karplus, K. (1985). "The Delta Rule Development System for Speech Synthesis from Text," Proc. IEEE 73, 1589–1601.
Hiki, S. (1970). "Control Rule of the Tongue Movement for Dynamic Analog Speech Synthesis," J. Acoust. Soc. Am. Suppl. 1 47, S85.
Hill, K., and Nessly, L. (1973). "Review of the Sound Pattern of English," Linguistics 106, 57–101.
Hirsh, I. J., Davis, H., Silverman, S.R., Reynolds, E.G., Eldest, E., and Benson, R. W. (1952). "Development of Materials for Speech Audiometry," J. Speech Hear. Disord. 17, 321–337.
Hirshberg, J., and Pierrehumbert, J. (1986). "The Intonational Structuring of Discourse," Proc. 24th Meeting Assoc. Comp. Linguistics, Columbia Univ., New York, pp. 136–144.
Hoard, J. E. (1966). "Juncture and Syllable Structure in English," Phonetica 15, 96–109.
Hoard, J. E. (1971). "Aspiration, Tenseness and Syllabification in English," Language 47, 133–140.
Hockett, C. (1955). *Manual of Phonetics* (Waverley, Baltimore, MD).
Holbrook, A., and Fairbanks, G. (1962). "Diphthong Formants and their Movements," J. Speech Hear. Res. 5, 38–58.
Holmes, J. N. (1961). "Research on Speech Synthesis Carried out during a Visit to the Royal Institute of Technology, Stockholm, from November 1960 to March 1961," Joint Speech Research Unit Report JU 11.4, British Post Office, Eastcote, England.
Holmes, J. N. (1973). "The Influence of the Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," IEEE Trans. Audio Electroacoust. AU-21, 298–305.
Holmes, J. N. (1983). "Formant Synthesizers: Cascade or Parallel," Speech Commun. 2, 251–273.
Holmes, J. N. (1984). "Speech Technology in the Next Decades," *Proceedings of the Tenth International Congress of Phonetic Sciences*, edited by M. P. R. Van den Broecke and A. Cohen (Foris, Dordrecht, The Netherlands), pp. 125–139.
Holmes, J. N., Mattingly, I. G., and Shearme, J. N. (1964). "Speech Synthesis by Rule," Lang. Speech 7, 127–143.
House, A. S. (1961). "On Vowel Duration in English," J. Acoust. Soc. Am. 33, 1174–1178.
House, A. S., and Fairbanks, G. (1953). "The Influence of Consonantal Environment upon the Secondary Acoustical Characteristics of Vowels," J. Acoust. Soc. Am. 25, 105–113.
House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (1965). "Articulation-Testing Methods: Consonant Differentiation with a Closed-Response Set," J. Acoust. Soc. Am. 37, 158–166.
Huggins, A. W. F., Houde, R. A., and Colwell, E. (1986). "Vidvox Human Factors Investigation," Bolt Beranek and Newman Inc. Report No. 6187, Cambridge, MA, p. 15.
Hunnicutt, S. (1976). "Phonological Rules for a Text-to-Speech System," Am. J. Comp. Ling. Microfiche 57, 1–72.
Hunnicutt, S. (1980). "Grapheme-to-Phoneme Rules, A Review," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR 2-3, 38–60.
Hunnicutt, S. (1984). "Bliss Symbol-to-Speech Conversion: Blisstalk," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR 1, 58–77.
Hunnicutt, S. (1985). "Lexical Prediction for a Text-to-Speech System," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden QPSR 2-3, 47–55.
IEEE (1969). "Recommended Practice for Speech Quality Measurements," IEEE Trans. Audio Electroacoust. AU-17, 225–246.
Ingemann, F. (1978). "Speech Synthesis by Rule using the FOVE Program," Status Report on Speech Research SR-54, Haskins Laboratories, New Haven, CT, 165–173.
Ishizaka, K., and Flanagan, J. L. (1972). "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," Bell Syst. Tech. J. 51, 1233–1268.
Ishizaka, K., and Matsudaira, M. (1968). "What Makes the Vocal Cords Vibrate?," in *The Sixth International Congress on Acoustics*, edited by Y. Kohasi (Elsevier, New York), Vol. II, pp. B9–B12.
Itakura, F., and Saito, S. (1968). "Analysis-Synthesis Telephony based on the Maximum Likelihood Method," Proc. 6th International Congress Acoustics, Tokyo, paper C-5-5.
Joos, M. A. (1948). "Acoustic Phonetics," Language Suppl. 24, 1–136.
Kahn, M. (1975). "Arabic Emphatics: The Evidence for Cultural Determinants of Phonetic Sex-Typing," Phonetica 31, 38–50.
Kalikow, D., Stevens, K. N., and Elliott, L. (1977). "Development of a Test of Speech Intelligibility in Noise Using Sentence Materials with Controlled Word Predictability," J. Acoust. Soc. Am. 61, 1337–1351.
Kaplan, G., and Lerner, E. J. (1985). "Realism in Synthetic Speech," IEEE Spectrum 22 (April), 32–37.
Kaplan, R., and Bresnan, J. (1982). "Lexical-Functional Grammar: A Formal System for Grammatical Representation," in *The Mental Representation of Grammatical Relations*, edited by J. Bresnan (MIT, Cambridge, MA).
Kelly, J., and Gerstman, L. (1961). "An Artificial Talker Driven from Phonetic Input," J. Acoust. Soc. Am. Suppl. 1 33, S35.
Kelly, J., and Gerstman, L. (1962). "Digital Computer Synthesizes Human Speech," Bell Labs. Rec. 40, 216–217.
Kelly, J., and Gerstman, L. (1964). "Synthesis of Speech from Code Signals," U.S. Patent 3,158,685.
Kelly, J., and Lochbaum, C. (1962). "Speech Synthesis," Proc. Fourth Int. Congr. Acoust. paper G42, 1–4 (Reprinted in Flanagan and Rabiner, 1973).
Kewley-Port, D. (1982). "Measurement of Formant Transitions in Naturally Produced Stop Consonant-Vowel Syllables," J. Acoust. Soc. Am. 72, 379–389.
Klatt, D. H. (1970). "Synthesis of Stop Consonants in Initial Position," J. Acoust. Soc. Am. Suppl. 1 47, S93.
Klatt, D. H. (1972). "Acoustic Theory of Terminal Analog Speech Synthesis," Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-72, 131–135.
Klatt, D. H. (1973a). "Durational Characteristics of Prestressed Word-Initial Consonant Clusters in English," Research Laboratory of Electronics QPR 108, MIT, Cambridge, MA, pp. 253–260.
Klatt, D. H. (1973b). "Interaction between Two Factors that Influence Vowel Duration," J. Acoust. Soc. Am. 54, 1102–1104.
Klatt, D. H. (1974). "The Duration of [S] in English Words," J. Speech Hear. Res. 17, 51–63.
Klatt, D. H. (1975a). "Vowel Lengthening is Syntactically Determined in a Connected Discourse," J. Phonetics 3, 129–140.
Klatt, D. H. (1975b). "Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters," J. Speech Hear. Res. 18, 686–706.
Klatt, D. H. (1976a). "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," J. Acoust. Soc. Am. 59, 208–1221.
Klatt, D. H. (1976b). "Structure of a Phonological Rule Component for a

Speech Synthesis by Rule Program," IEEE Trans. Acoust. Speech Signal Process. **ASSP-24**, 391–398.
Klatt, D. H. (**1979a**). "Synthesis by Rule of Segmental Durations in English Sentences," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, New York), pp. 287–300.
Klatt, D. H. (**1979b**). "Synthesis by Rule of Consonant-Vowel Syllables," Speech Communication Group Working Papers 3, MIT, Cambridge, MA, pp. 93–104.
Klatt, D. H. (**1980**). "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
Klatt, D. H. (**1981**). "A Text-to-Speech Conversion System," Proc. AFIPS Office Automation Conference, pp. 51–61.
Klatt, D. H. (**1982a**). "The Klattalk Text-to-Speech System," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-82**, 1589–1592.
Klatt, D. H. (**1982b**). "A Strategy for the Perceptual Interpretation of Durational Cues," Speech Communication Group Working Papers 1, MIT, Cambridge, MA, pp. 83–91.
Klatt, D. H. (**1982c**). "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-82**, 1278–1281.
Klatt, D. H. (**1986a**). "Representation of the First Formant in Speech Recognition and in Models of the Auditory Periphery," in Proc. Montreal Satellite Symposium on Speech Recognition, edited by P. Mermelstein, Twelfth Int. Cong. Acoustics, Toronto, Canada, pp. 5–7.
Klatt, D. H. (**1986b**). "Detailed Spectral Analysis of a Female Voice," J. Acoust. Soc. Am. Suppl. 1 **80**, S97.
Klatt, D. H. (**1987**). "How Klattalk became DECtalk: An Academic's Experiences in the Business World," Speech Tech. **87**, 293–294.
Klatt, D. H., and Aoki, C. (**1984**). "Synthesis by Rule of Japanese," J. Acoust. Soc. Am. Suppl. 1 **76**, S2.
Klatt, D. H., and Shipman, D. W. (**1982**). "Letter-to-Phoneme Rules: A Semi-Automatic Discovery Procedure," J. Acoust. Soc. Am. Suppl. 1 **72**, S48.
Koenig, W. H., Dunn, H. K., and Lacey, L. Y. (**1946**). "The Sound Spectrograph," J. Acoust. Soc. Am. **18**, 19–49.
Kučera, H., and Francis, W. N. (**1967**). *Computational Analysis of Present day American English* (Brown U.P., Providence, RI).
Kurzweil, R. (**1976**). "The Kurzweil Reading Machine: A Technical Overview," in *Science, Technology and the Handicapped*, edited by M. R. Redden and W. Schwandt (American Association for the Advancement of Science, Report 76-R-11, Washington, DC), pp. 3–11.
Labov, W. (**1986**). "Sources of Inherent Variation in the Speech Process," in *Invariance and Variability in Speech Processes*, edited by J. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 402–425.
Ladd, D. R. (**1983**). "Phonological Features of Intonational Peaks," Language **59**, 721–759.
Ladefoged, P. (**1973**). "The Features of the Larynx," J. Phonetics **1**, 73–83.
Lamel, L., and Zue, V. (**1984**). "Properties of Consonant Sequences within Words and across Word Boundaries," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-84**, 42.3.1–43.2.4.
Lawrence, W. (**1953**). "The Synthesis of Speech from Signals which have a Low Information Rate," in *Communication Theory*, edited by W. Jackson (Butterworths, London, England), pp. 460–469.
Lee, D., and Lochovsky, F. (**1983**). "Voice Response Systems," ACM Computing Surveys **15**, 351–374.
Lee, F. F. (**1969**). "Reading Machine: From Text to Speech," IEEE Trans. Audio Electroacoust. **AU-17**, 275–282.
Lehiste, I. (**1959**). "An Acoustic-Phonetic Study of Internal Open Juncture," Suppl. to Phonetica **5**, 1–55.
Lehiste, I. (**1962**). "Acoustical Characteristics of Selected English Consonants," Univ. Michigan Speech Research Lab. Report **9**, 1–219.
Lehiste, I. (**1964**). "Juncture," in *Proceedings of the Fifth International Congress of Phonetic Sciences*, edited by E. Zwirner and W. Bethge (Karger, Basal, Switzerland), pp. 172–200.
Lehiste, I. (**1967**). *Readings in Acoustic Phonetics* (MIT Press, Cambridge, MA).
Lehiste, I. (**1970**). *Suprasegmentals* (MIT Press, Cambridge, MA).
Lehiste, I. (**1975a**). "Some Factors Affecting the Duration of Syllabic Nuclei in English," *Proceedings of the First Salzburg Conference on Linguistics*, edited by G. Drachman (Verlag Gunter, Narr), pp. 81–104.
Lehiste, I. (**1975b**). "The Phonetic Structure of Paragraphs," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. Nooteboom (Springer, Heidelberg, Germany), pp. 195–206.
Lehiste, I. (**1977**). "Isochrony Reconsidered," J. Phonetics **5**, 253–263.
Lehiste, I. Olive, J. P., and Streeter, L. A. (**1976**). "The Role of Duration in Disambiguating Syntactically Ambiguous Sentences," J. Acoust. Soc. Am. **60**, 1199–1202.
Lehiste, I., and Peterson, G. E. (**1959**). "Linguistic Considerations in the Study of Speech Intelligibility," J. Acoust. Soc. Am. **31**, 280–287.
Lehiste, I., and Peterson, G. E. (**1961**). "Some Basic Considerations in the Analysis of Intonation," J. Acoust. Soc. Am. **33**, 419–425.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the Speech Code," Psychol. Rev. **74**, 431–461.
Liberman, A. M., Delattre, P., and Cooper, F. S. (**1958**). "Some Cues for the Distinction between Voiced and Voiceless Stops in Initial Position," Lang. Speech **1**, 153–167.
Liberman, A. M., Delattre, P., Cooper, F. S., and Gerstman, L. J. (**1954**). "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants," Psychol. Monogr. **68**, 1–13.
Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. (**1959**). "Minimal Rules for Synthesizing Speech," J. Acoust. Soc. Am. **31**, 1490–1499.
Liberman, A. M., and Mattingly, I. G. (**1985**). "The Motor Theory of Speech Perception Revisited," Cognition **21**, 1–36.
Liberman, M. Y. (**1979**). "Phonemic Transcription, Stress, and Segment Durations for Spelled Proper Names," J. Acoust. Soc. Am. Suppl. 1 **64**, S163.
Lieberman, P. (**1967**). *Intonation, Perception and Language* (MIT Press, Cambridge, MA).
Liljencrants, J. (**1969**). "Speech Synthesizer Control by Smoothed Step Functions," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden **QPSR-4**, 43–50.
Liljencrants, J. (**1985**). "Speech Synthesis with a Reflection-Type Line Analog," unpublished Ph.D. thesis, Dept. Speech Commun. and Musical Acoust., Royal Inst. of Tech., Stockholm, Sweden.
Lindblom, B. (**1963**). "Spectrographic Study of Vowel Reduction," J. Acoust. Soc. Am. **35**, 1773–1781.
Lisker, L. (**1957**). "Minimal Cues for Separating /wrly/ in Intervocalic Position," Word **13**, 256–267.
Lisker, L. (**1978**). "Rapid vs. Rabid: A Catalog of Acoustic Features that may Cue the Distinction," Status Report on Speech Research 54, Haskins Laboratories, New Haven, CT, pp. 127–132.
Lisker, L., and Abramson, A. S. (**1967**). "Some Effects of Context on Voice Onset Time in English Stops," Lang. Speech **10**, 1–28.
Logan, J. S., and Pisoni, D. B. (**1986**). "Preference Judgements Comparing Different Synthetic Voices," J. Acoust. Soc. Am. Suppl. 1 **79**, S24.
Logan, J. S., Pisoni, D. B., and Greene, B. G. (**1986**). "Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to-Speech Systems," submitted to J. Acoust. Soc. Am.
Lucassen, J. M., and Mercer, R. L. (**1984**). "An Information Theoretic Approach to the Automatic Determination of Phonemic Base Forms," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-84**, 42.5.1–42.5.4.
Luce, P., Feustel, T., and Pisoni, D. (**1983**). "Capacity Demands in Short-Term Memory for Synthetic and Natural Speech," Human Factors **25**, 17–31.
MacNeilage, P. F., and DeClerk, J. L. (**1969**). "On the Motor Control of Coarticulation of CVC Syllables," J. Acoust. Soc. Am. **45**, 1217–1233.
Maeda, S. (**1974**). "A Characterization of Fundamental Frequency Contours of Speech," Research Laboratory of Electronics QPR 114, MIT, Cambridge, MA, pp. 193–211.
Maeda, S. (**1987**). "On the Generation of Sounds in Stop Consonants," Speech Communication Group Working Papers V, MIT, Cambridge, MA, pp. 1–14.
Magnusson, L., Blomberg, M., Carlson, R., Elenius, K., and Granström, B. (**1984**). "Swedish Speech Researchers Team Up with Electronic Venture Capitalists," Speech Technol. **2**, 15–24.
Makhoul, J. (**1973**). "Spectral Analysis of Speech by Linear Prediction," IEEE Trans. Audio Electroacoust. **AU-21**, 140–148.
Malecot, A. (**1956**). "Acoustic Cues for Nasal Consonants: An Experimental Study Involving a Tape-Splicing Technique," Language **32**, 274–284.
Malme, C. I. (**1959**). "Detectability of Small Irregularities in a Broadband Noise Spectrum," Research Lab. of Electronics Q.P.R. 52, Mass. Inst. Tech., pp. 139–141.
Manous, L. M., Pisoni, D. B., Dedina, M. J., and Nusbaum, H.C. (**1985**). "Comprehension of Natural and Synthetic Speech Using a Sentence Verification Task," Speech Research Laboratory Progress Report 11, In-

diana University, Bloomington, IN, pp. 33–58.
Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language* (MIT Press, Cambridge, MA).
Markel, J. D. (1972). "Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation," IEEE Trans. Audio Electroacoust. **AU-20**, pp. 129–137.
Martin, J. G., and Bunnell, H. T. (1982). "Perception of Anticipatory Coarticulation Effects in Vowel-Stop Consonant-Vowel Sequences," J. Exp. Psychol.: Human Percept. Perform. **8**, 473–488.
Mathews, M. V., Miller, J. E., and David, E. E. (1961). "Pitch-Synchronous Analysis of Voiced Sounds," J. Acoust. Soc. Am. **33**, 179–186.
Matsui, E., Suzuki, T., Umeda, N., and Omura, H. (1968). "Synthesis of Fairy Tales using an Analog Vocal Tract," Proc. 6th International Congress on Acoustics, Tokyo, Japan, B159–162.
Mattingly, I. G. (1966). "Synthesis by Rule of Prosodic Features," Lang. Speech **9**, 1–13.
Mattingly, I.G. (1968). "Synthesis-by-Rule of General American English," Supplement to Status Report on Speech Research, Haskins Laboratories, New Haven, CT, pp. 1–223.
Mattingly, I. G. (1974). "Speech Synthesis for Phonetic and Phonological Models," in *Current Trends in Linguistics*, edited by T. S. Sebeok, Mouton, 's-Gravenhague, The Netherlands), Vol. 12, pp. 2451–2487.
Maxey, H. D. (1963). "Terminal Analog Synthesis of Voiced Fricatives," J. Acoust. Soc. Am. **35**, 1890 (A).
McIlroy, M. D. (1974). "Synthetic Speech by Rule," unpublished Technical Memo, Bell Laboratories, Murray Hill, NJ.
McGillivray, R. (1983). "Terminals with Voice Output," in *Aids and Appliances Review 9* (Carroll Center for the Blind, 770 Centre St., Newton, MA 02158), pp. 32–41.
Mermelstein, P. (1973). "Articulatory Model for the Study of Speech Production," J. Acoust. Soc. Am. **53**, 1070–1082.
Miller, R. L. (1959). "Nature of the Vocal Cord Wave," J. Acoust. Soc. Am. **31**, 667–677.
Minow, M., and Klatt, D. H. (1983). "DECtalk DTC01 Text-to-Speech System Owner's Manual," EK-DTC01-OM-001, Educational Services of Digital Equipment Corporation, P.O. Box CS2008, Nashua, NH 03061.
Monsen, R. B., and Engebretson, A. M. (1977). "Study of Variation in the Male and Female Glottal Wave," J. Acoust. Soc. Am. **62**, 981–993.
Morris, L. R. (1979). "A Fast Fortran Implementation of the NRL Algorithm for Automatic Translation of English Text to Votrax Parameters," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-79**, 907–913.
Munson, W., and Karlin, J. (1962). "Isopreference Method for Evaluating Speech Transmission Circuits," J. Acoust. Soc. Am. **27**, 1213–1219.
Nakata, K., and Mitsuoka, T. (1965). "Phonemic Transformation and Control Aspects of Synthesis of Connected Speech," J. Radio Res. Labs. **12**, 171–186.
Nakatani, L. H., and Dukes, K. D. (1973). "Sensitive Test of Speech Communication Quality," J. Acoust. Soc. Am. **53**, 1083–1092.
Nakatani, L. H., and Dukes, K. D. (1977). "Locus of Segmental Cues for Word Juncture," J. Acoust. Soc. Am. **62**, 714–719.
Nakatani, L. H., Egan, D., Ruedisueli, L., and Hawley, P. (1986). "A Talking Tutor and Trainer for Teaching the Use of Interactive Computer Systems," Proc. Conf. on Human Factors in Computing Systems (ACM, New York), pp. 29–34.
Nakatani, L., and O'Connor-Dukes, K. (1979). "Phonetic Parsing Cues for Word Perception," unpublished Technical Memorandum, Bell Laboratories, Murray Hill, NJ.
Nakatani, L. H., and Schaffer, J. (1978). "Hearing Words without Words: Prosodic Cues for Word Perception," J. Acoust. Soc. Am. **63**, 234–244.
NCHS (1977). "Pamphlet from the National Center for Health Statistics," Washington, DC.
Nixon, C. W., Anderson, T. R., and Moore, T. J. (1985). "The Perception of Synthetic Speech in Noise," Armstrong Aerospace Medical Research Laboratories Report, Wright-Patterson AFB, OH 45433.
Nusbaum, H. C., Dedina, M. J., and Pisoni, D. B. (1984). "Perceptual Confusions of Consonants in Natural and Synthetic CV Syllables," Speech Research Laboratory Progress Report 10, Indiana Univ., Bloomington, IN, pp. 409–422.
Nusbaum, H. C., Pisoni, D. B., and Schwab, E. C. (1984). "Subjective Evaluation of Synthetic Speech: Measuring Preference, Naturalness and Intelligibility," Speech Research Laboratory Progress Report 10, Indiana Univ., Bloomington, IN, pp. 391–408.
Nye, P., and Gaitenby, J. (1974). "The Intelligibility of Synthetic Monosyllable Words in Short, Syntactically Normal Sentences," Status Report on Speech Research SR-37/38, Haskins Laboratories, New Haven, CT, pp. 169–190.
Nye, P., Hankins, J., Rand, T., Mattingly, I., and Cooper, F. (1973). "A Plan for the Field Evaluation of an Automated Reading System for the Blind," IEEE Trans. Audio Electroacoust. **AU-21**, 265–268.
Ochsman, R. B., and Chapanis, A. (1974). "The Effects of 10 Communication Modes on the Behavior of Teams during Cooperative Problem Solving," Int. J. Man-Machine Studies **6**, 579–619.
O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. G., and Cooper, F. S. (1957). "Acoustic Cues for the Perception of Initial /w,j,r,l/ in English," Word **13**, 24–36.
Öhman, S. E. G. (1966). "Coarticulation in VCV Utterances: Spectrographic Measurements," J. Acoust. Soc. Am. **39**, 151–168.
Öhman, S. E. G. (1967). "Word and Sentence Intonation: A Quantitative Model," Speech Transmission Laboratory, Royal Institute of Technology, Stockholm **QPSR 2-3**, 20–54.
Olive, J. P. (1977). "Rule Synthesis of Speech from Diadic Units," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-77**, 568–570.
Olive, J. P., and Liberman, M. Y. (1979). "A Set of Concatenative Units for Speech Synthesis," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, edited by J J. Wolf and D. H. Klatt (American Institute of Physics, New York), pp. 515–518.
Olive, J. P., and Liberman, M. Y. (1985). "Text-to-Speech—An Overview," J. Acoust. Soc. Am. Suppl. 1 **78**, S6.
Olive, J. P., and Nakatani, L. H. (1974). "Rule Synthesis of Speech by Word Concatenation: A First Step," J. Acoust. Soc. Am. **55**, 660–666.
Olive, J. P., and Spickenagle, N. (1976). "Speech Resynthesis from Phoneme-Related Parameters," J. Acoust. Soc. Am. **59**, 993–996.
Oller, D. K. (1973). "The Effect of Position in Utterance on Speech Segment Duration in English," J. Acoust. Soc. Am. **54**, 1235–1247.
Olson, R., Foltz, G., and Wise, B. (1985). "Reading Instruction and Remediation with the Aid of Computer Speech," oral presentation, Meeting of the Society for Computers in Psychology, Boston, MA.
O'Malley, M. H., Larkin, D. K., and Peters, E. W. (1986). "Beyond the Reading Machine: What the Next Generation of Intelligent Text-to-Speech Systems Should Do for the User," in *Proceedings Speech Tech '86* (Media Dimensions, New York), pp. 216–219.
O'Shaughnessy, D. (1977). "Fundamental Frequency by Rule for a Text-to-Speech System," Proc. Int. Conf. Acoust. Speech Signal Process. **ICASSP-77**, 571–574.
O'Shaughnessy, D. (1979). "Linguistic Features in Fundamental Frequency Patterns," J. Phonetics **7**, 119–145.
O'Shaughnessy, D., and Allen, J. (1983). "Linguistic Modality Effects on Fundamental Frequency in Speech," J. Acoust. Soc. Am. **74**, 1155–1171.
Oshika, B., Zue, V., Weeks, R., Neu, H., and Aurbach, J. (1975). "The Role of Phonological Rules in Speech Understanding Research," IEEE Trans. Acoust. Speech Signal Proc. **ASSP-23**, 104–110.
Peck, C. (1969). "An Acoustic Investigation of the Intonation of American English," Nat. Lang. Studies I, Univ. Michigan, Ann Arbor, MI, pp. 36–47.
Peterson, G. E., and Barney, H. L. (1952). "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. **24**, 175–184.
Peterson, G. E., and Lehiste, I. (1960). "Duration of Syllabic Nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.
Peterson, G., Wang, W., and Sivertsen, E. (1958). "Segmentation Techniques in Speech Synthesis," J. Acoust. Soc. Am. **30**, 739–742.
Pierrehumbert, J. (1981). "Synthesizing Intonation," J. Acoust. Soc. Am. **70**, 985–995.
Pike, K. L. (1945). *The Intonation of American English* (Univ. Michigan Press, Ann Arbor, MI).
Pisoni, D. B. (1982). "Perception of Speech: The Human Listener as a Cognitive Interface," Speech Technol. **1**, 10–23.
Pisoni, D. B., and Hunnicutt, S. (1980). "Perceptual Evaluation of MITalk: The MIT Unrestricted Text-to-Speech System," Proc Int. Conf. Acoust. Speech Signal Process. **ICASSP-80**, 572–575.
Pisoni, D. B., and Koen, E. (1981). "Some Comparisons of Intelligibility of Synthetic and Natural Speech at Different Speech-to-Noise Ratios," Speech Research Laboratory Progress Report 7, Indiana Univ., Bloomington, IN, pp. 243–254.
Pisoni, D. B., Nusbaum, H. C., and Greene, B. G. (1985). "Perception of Synthetic Speech Generated by Rule," Proc. IEEE **73**, 1665–1676.
Pols, L. C. W., and Olive, J. P. (1983). "Intelligibility of Consonants in CVC Utterances Produced by Diadic Rule Synthesis," Speech Commun.

2, 3–13.
Portnoy, K. (Ed.) (1979–present). *Communication Outlook* (Artificial Language Lab, Michigan State Univ., Lansing, MI).
Potter, R. K. (1946). "Introduction to Technical Discussions of Sound Portrayal," J. Acoust. Soc. Am. 18, 1–3. (See also five related articles that follow.)
Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech* (van Nostrand, New York).
Quarmby, D. J., and Holmes, J. N. (1984). "Implementation of a Parallel-Formant Speech Synthesizer using a Single-Chip Programmable Signal Processor," IEE Proceed. 131, 563–569.
Rabiner, L. R. (1968). "Speech Synthesis by Rule: An Acoustic Domain Approach," Bell Syst. Tech. J. 47, 17–38.
Rabiner, L. R. (1969). Private communication.
Rabiner, L. R., Schafer, R. W., and Flanagan, J. L. (1971). "Computer Synthesis of Speech by Concatenation of Formant-Coded Words," Bell System Tech. J. 50, 1541–1558.
Repp, B. H. (1986). "Perception of the [m]-[n] Distinction in CV Syllables," J. Acoust. Soc. Am. 79, 1987–1999.
Rodet, X. (1984). "Time-Domain Formant-Wave-Function Synthesis," Comput. Music J. 8, 9–14.
Rosen, G. (1958). "A Dynamic Analog Speech Synthesizer," J. Acoust. Soc. Am. 30, 201–209.
Rosenberg, A. (1971). "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," J. Acoust. Soc. Am. 49, 583–590.
Rothenberg, M., Carlson, R., Granström, B., and Gauffin, J. (1975). "A Three-Parameter Voice Source for Speech Synthesis," in *Speech Communication*, edited by G. Fant (Almqvist and Wiksell, Uppsala, Sweden), Vol. 2, pp. 235–243.
Schwab, E. C., Nusbaum, H. C., and Pisoni, D. B. (1986). "Effects of Training on the Perception of Synthetic Speech," Human Factors 37, 395–408.
Schwartz, M. F. (1967). "Transitions in American English /s/ as Cues to the Identity of Adjacent Stop Consonants," J. Acoust. Soc. Am. 42, 897–899.
Sejnowski, T. J., and Rosenberg, C. R. (1986). "NETtalk: A Parallel Network that Learns to Read Aloud," Elec. Engr. Comp. Sci. Tech. Report JHU/EECS-86/01, Johns Hopkins University, Baltimore, MD, pp. 1–23.
Shearme, J. N., and Holmes, J. N. (1961). "An Experimental Study of the Classification of Sounds in Continuous Speech According to their Distribution in the Formant 1–Formant 2 Plane," Proc. Fourth Int. Cong. Phonetic Sci., pp. 234–240.
Sherwood, B. A. (1979). "The Computer Speaks," IEEE Spectrum, August, 18–25.
Sherwood, B. A. (1981). "New Technology Provides Computer Voices for Education," Speech Technol. 1, 24–29.
Shoup, J. E. (1980). "Phonological Aspects of Speech Recognition," in *Trends in Speech Recognition*, edited by W. A. Lea (Prentice-Hall, New York), pp. 125–138.
Sivertsen, E. (1961). "Segment Inventories for Speech Synthesis," Lang. Speech 4, 27.
Social Security Administration (1985). *Report of Distribution of Surnames in the Social Security Number File, September 1, 1984* (Department of Health and Human Resources, Washington, DC), Social Security Administration Pub. No. 42-004.
Stevens, K. N. (1972). "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in *Human Communication: A Unified View*, edited by E. E. David and P. B. Denes (McGraw-Hill, New York), pp. 51–66.
Stevens, K. N. (1977). "Physics of Larynx Behavior and Larynx Modes," Phonetica 34, 264–279.
Stevens, K. N. (1987). Book in preparation.
Stevens, K. N., Fant, G., and Hawkins, S. (1987). "Some Acoustical and Perceptual Correlates of Nasal Vowels," in *Festshrift for Ilse Lehiste*, edited by R. Channon and L. Shockey (Foris, Dordrecht, The Netherlands), pp. 241–254.
Stevens, K. N., and House, A. S. (1955). "Development of a Quantitative Description of Vowel Articulation," J. Acoust. Soc. Am. 27, 484–493.
Stevens, K. N., and House, A. S. (1961). "An Acoustical Theory of Vowel Production and Some of Its Implications," J. Speech Hear. Res. 4, 303–320.
Stevens, K. N., and House, A. S. (1963). "Perturbation of Vowel Articulation by Consonantal Context: An Acoustical Study," J. Speech Hear. Res. 6, 111–128.
Stevens, K. N., Kasowski, S., and Fant, G. (1953). "An Electrical Analog of the Vocal Tract," J. Acoust. Soc. Am. 25, 734–742.
Stewart, J. Q. (1922). "An Electrical Analogue of the Vocal Organs," Nature 110, 311–312.
Streeter, L. (1978). "Acoustic Determinants of Phrase Boundary Perception," J. Acoust. Soc. Am. 64, 1582–1592.
Sundberg, J., and Gauffin, J. (1979). "Waveform and Spectrum of Glottal Voice Source," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Ohman (Academic, London) pp. 301–322.
Suppes, P. (1979). "Current Trends in Computer Assisted Instruction," in *Advances in Computers* (Academic, New York).
Suppes, P. (Ed.) (1981). *University-Level Computer-Assisted Instruction at Stanford: 1968–1980* (Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, CA).
Svensson, S. G. (1974). "Prosody and Grammar in Speech Perception," *Monographs from the Institute of Linguistics* (University of Stockholm, Stockholm, Sweden).
Teranishsi, R., and Umeda, N. (1968). "Use of Pronouncing Dictionary in Speech Synthesis Experiments," Reports of the Sixth International Congress on Acoustics, Tokyo, Vol. II, pp. 155–158.
Titze, I. R. (1974). "The Human Vocal Cords: a Mathematical Model," Phonetica 29, 1–21.
Titze, I. R., and Talkin, D. (1979). "A Theoretical Study of the Effects of the Various Laryngeal Configurations on the Acoustics of Phonation," J. Acoust. Soc. Am. 66, 60–74.
Uldall, E. (1960). "Attitudinal Meanings Conveyed by Intonation Contours," Lang. Speech 3, 223–234.
Umeda, N. (1975). "Vowel Duration in American English," J. Acoust. Soc. Am. 58, 434–445.
Umeda, N. (1976). "Linguistic Rules for Text-to-Speech Synthesis," Proc. IEEE 64, 443–451.
Umeda, N. (1977). "Consonant Duration in American English," J. Acoust. Soc. Am. 61, 846–858.
Umeda, N. (1981). "Influence of Segmental Factors on Fundamental Frequency in Fluent Speech," J. Acoust. Soc. Am. 70, 350–355.
Umeda, N., and Coker, C. H. (1974). "Allophonic Variation in American English," J. Phonetics 2, 1–5.
Umeda, N., Matsui, E., Suzuki, T., and Omura, H. (1968). "Synthesis of Fairy Tales using an Analog Vocal Tract," Proc. 6th International Congress on Acoustics, Tokyo, Japan, pp. B159–162.
Umeda, N., and Quinn, A. M. S. (1981). "Word Duration as an Acoustic Measure of Boundary Perception," J. Phonetics 9, 19–28.
Umeda, N., and Teranishi, R. (1975). "The Parsing Program for Automatic Text-to-Speech Synthesis Developed at the Electrotechnical Laboratory in 1968," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 183–188.
Vanderheiden, G. C. (1978). *Nonvocal Communication Resource Book* (University Park, Baltimore, MD).
Vanderheiden, G. C. (1985). "1985 Updates for the Nonvocal Communication Resource Book," Trace Center of the Waisman Center, University of Wisconsin, Madison, WI.
Vanderslice, R. (1968). "Synthetic Elocution," Working Papers in Phonetics, UCLA, Los Angeles, CA 8, 1–131.
Vanderslice, R., and Ladefoged, P. (1971). "Binary Suprasegmental Features and Transformational Word-Accentuation Rules," Language 48, 819–839.
Venezky, R. L. (1965). *A Study of English Spelling-to-Sound Correspondences on Historical Principles* (Ann Arbor Press, Ann Arbor, MI).
Venezky, R. L. (1970). *The Structure of English Orthography* (Mouton,'s-Gravenhague, The Netherlands).
Voiers, W. (1983). "Evaluating Processed Speech using the Diagnostic Rhyme Test," Speech Technol. 1, 30–39.
Wang, W., and Peterson, G. E. (1958). "Segment Inventory for Speech Synthesis," J. Acoust. Soc. Am. 30, 743–746.
Warrick, A., Nelson, P. J., Cossaltor, J. G., Cote, C., McGillis, J., and Charbonneau, J. R. (1977). "Synthesized Speech as an Aid to Communication and Learning for the Non-Verbal," Proc. Workshop on Communication Aids for the Handicapped, Ottawa, Canada, pp. 120–135.
Werner, E., and Haggard, M. (1969). "Articulatory Synthesis by Rule," in Speech Synthesis and Perception (Psychological Laboratory, Univ. Cambridge, Cambridge, England), Vol. I, pp. 1–35.
Wiggins, R. (1980). "An Integrated Circuit for Speech Synthesis," IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP-80, 398–401.

Wijk, A. (1969). In *Alphabets for English*, edited by W. Haas (Manchester Univ. Press, Manchester, England).
Woods, W. A. (1970). "Transition Network Grammars for Natural Language Analysis," Commun. ACM **13**, 591–606.
Wright, C. E., Altom, M. J., and Olive, J. P. (1986). "Diagnostic Evaluation of a Synthesizer's Acoustic Inventory," J. Acoust. Soc. Am. Suppl. 1 **79**, S25.
Wright, J. T., Malsheen, B. J., and Peet, M. (1986). "Comparison of Segmental Intelligibility and Pronunciation Accuracy for Two Commercial Text-to-Speech Systems," Proc. Am. Voice Input/Output Soc. AVIOS-86, 235–261.
Young, R. W. (1948). "Review of U.S. Patent 2,432,321, Translation of Visual Symbols, R. K. Potter, assignor (9 December 1947)," J. Acoust. Soc. Am. **20**, 888–889.
Young, S. J., and Fallside, F. (1979). "Speech Synthesis from Concept: A Method for Speech Output from Information Systems," J. Acoust. Soc. Am. **66**, 685–695.
Zue, V. W., and Laferriere, M. (1979). "An Acoustic Study of Medial /t,d/ in American English," J. Acoust. Soc. Am. **66**, 1039–1050.
Zue, V. W., and Shattuck-Hufnagel, S. (1979). "The Palatalization of Alveolar Fricatives in American English," Proc. 9th Int. Cong. Phonetics, Copenhagen, Denmark, pp. 215–216.