Review of text-to-speech conversion for English

Dennis H. Klatt

Citation: The Journal of the Acoustical Society of America **82**, 737 (1987); doi: 10.1121/1.395275
View online: https://doi.org/10.1121/1.395275
View Table of Contents: https://asa.scitation.org/toc/jas/82/3
Published by the Acoustical Society of America

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

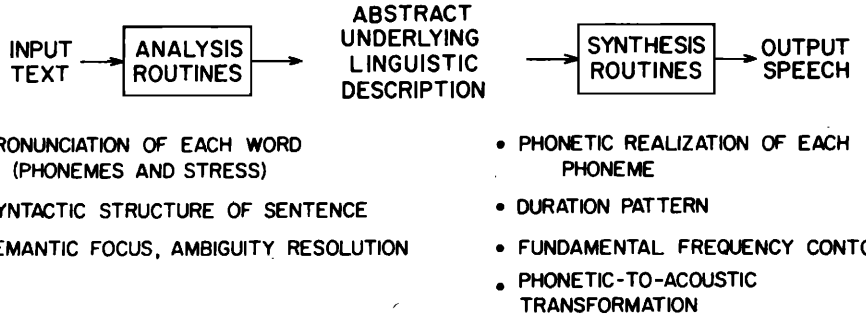


FIG. 1. Text must be converted to an abstract linguistic representation so as to be able to generate an accurate synthetic approximation to an English sentence.

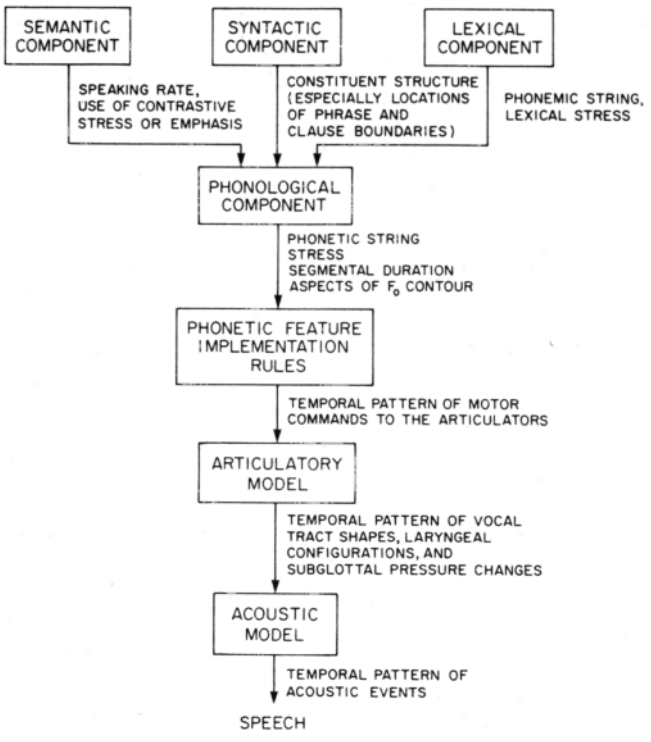


FIG. 2. Simplified block diagram of how a linguist might view the sentence generation process. An abstract linguistic representation for a sentence that is provided by the semantic component, syntactic component and lexical component undergoes various intermediate transformations before becoming an acoustic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

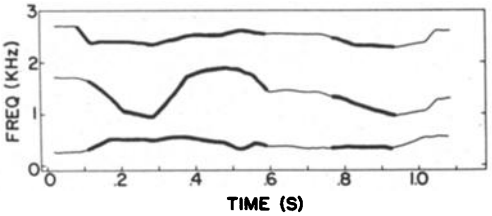
ABSTRACT LINGUISTIC REPRESENTATION:
/ǰo)'et hɪz s'up./

ALLOPHONIC RECODING:
[ǰo 'eɾ ɪs s'up.]

DURATION SPECIFICATION, IN MSEC:
[100, 210, 180, 20, 65, 75, 90, 165, 75]

FUNDAMENTAL FREQUENCY GESTURES:
1. HAT RISE DURING [o]
2. STRESS PULSE ON [o]
3. STRESS PULSE ON [e]
4. STRESS PULSE ON [u]
5. HAT FALL DURING [u]

SELECTED SYNTHESIZER CONTROL PARAMETERS:



SPECTROGRAM OF SYNTHETIC WAVEFORM:

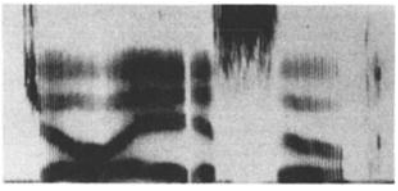


FIG. 3. An example of successive stages in the Klattalk transformation of a phonemic representation for the sentence "Joe ate his soup" to an acoustic waveform, see text.

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

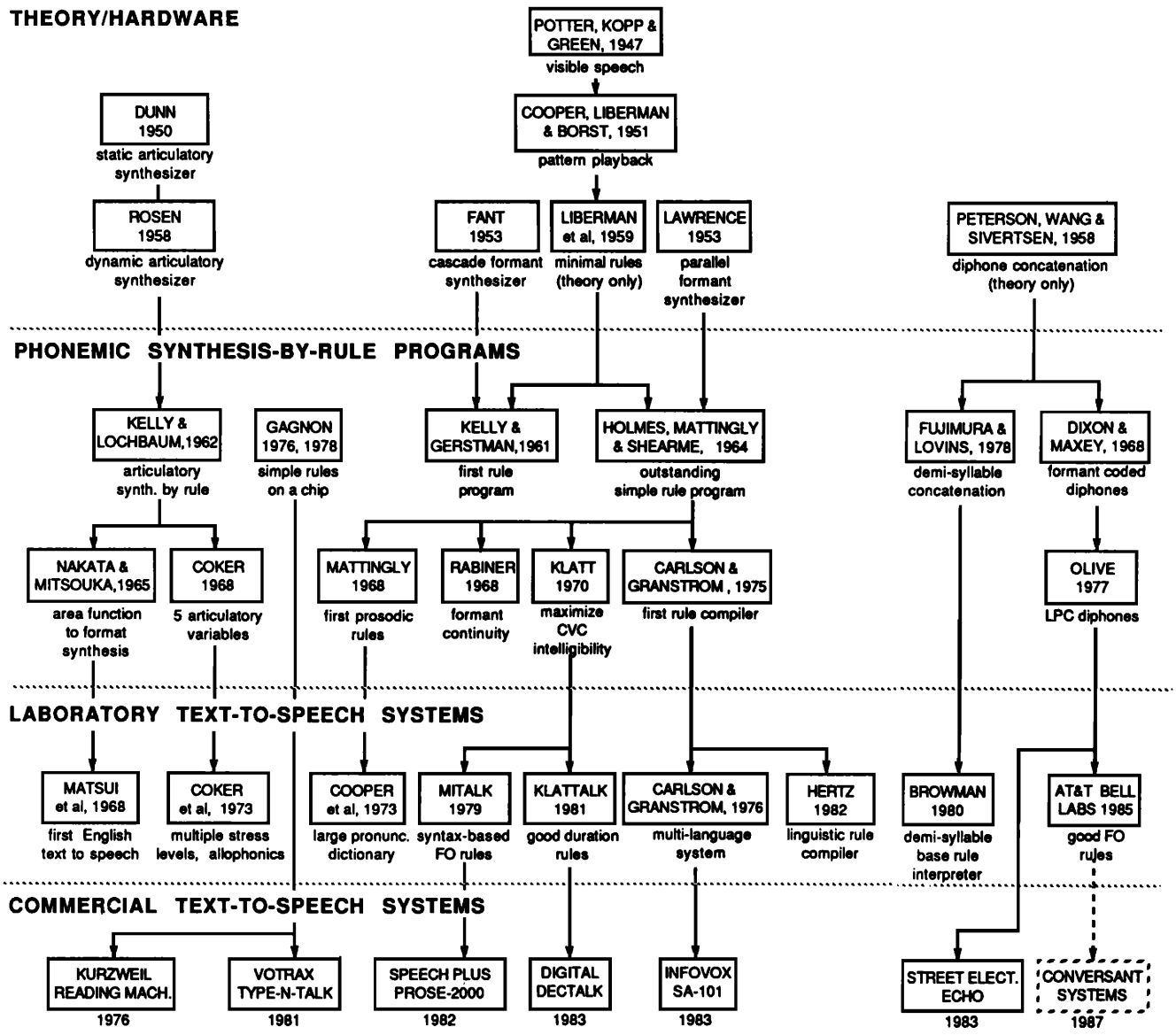


FIG. 4. Historical antecedents of the phoneme-to-speech algorithms used in several commercial text-to-speech systems.

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-

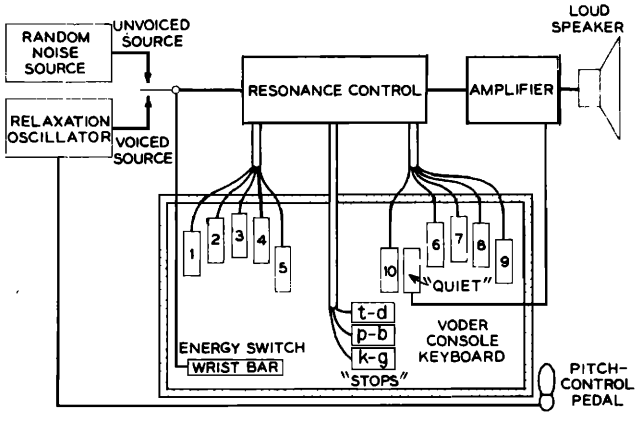


FIG. 5. The Voder speech synthesizer, consisting of a bank of filters excited by an impulse train or noise, and controlled by a piano-like keyboard, after Dudley *et al.* (1939).

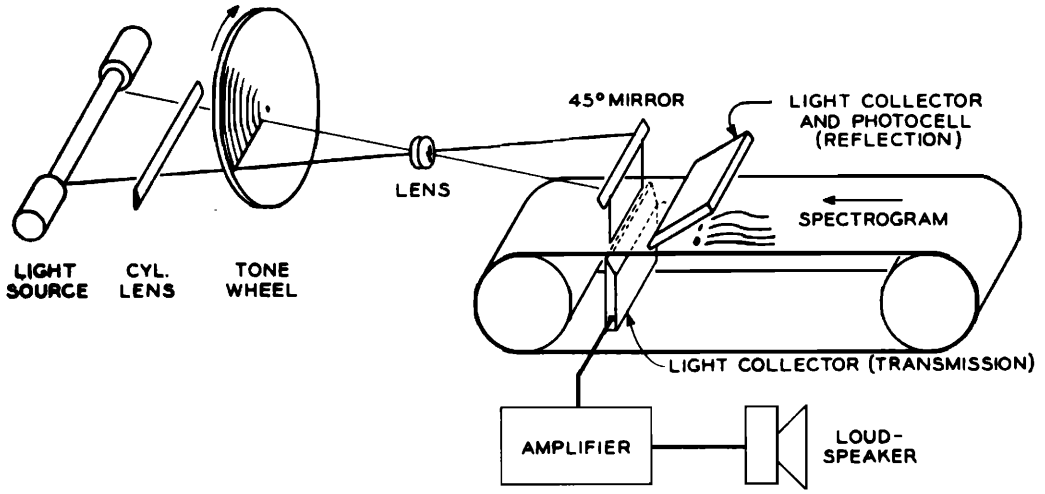


FIG. 6. The Haskins Pattern Playback, consisting of an optical system for modulating the amplitudes of a set of harmonics of 120 Hz over time depending on patterns painted on a moving transparent belt, after Cooper *et al.* (1951).

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-

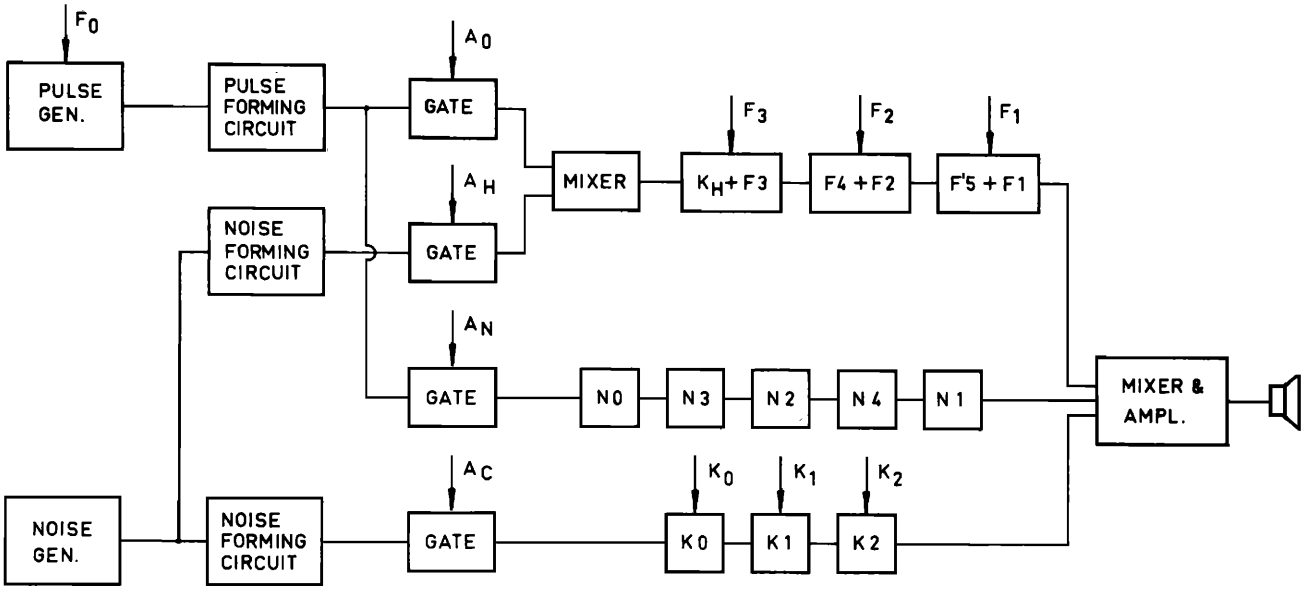


FIG. 7. The OVE II speech synthesizer, consisting of three separate circuits to model the transfer function of the vocal tract for vowels (top), nasals (middle), and obstruent consonants (bottom), after Fant and Martony (1962). Available sound sources are voicing (top), aspiration noise (middle), and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

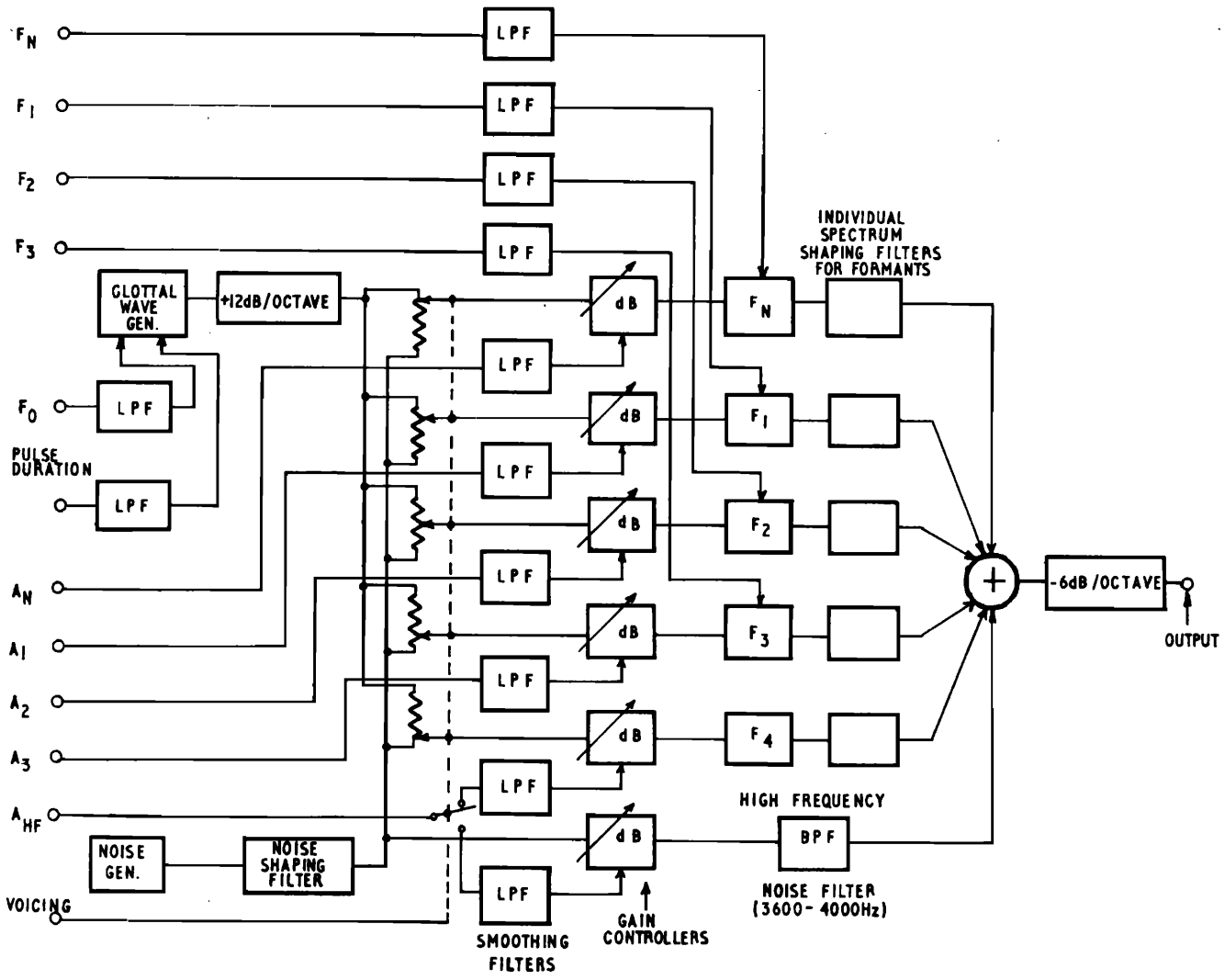


FIG. 8. The Holmes parallel formant synthesizer, consisting of four parallel formants and a nasal formant, each excited by a variable mixture of voicing and/or noise, after Holmes (1973).

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-

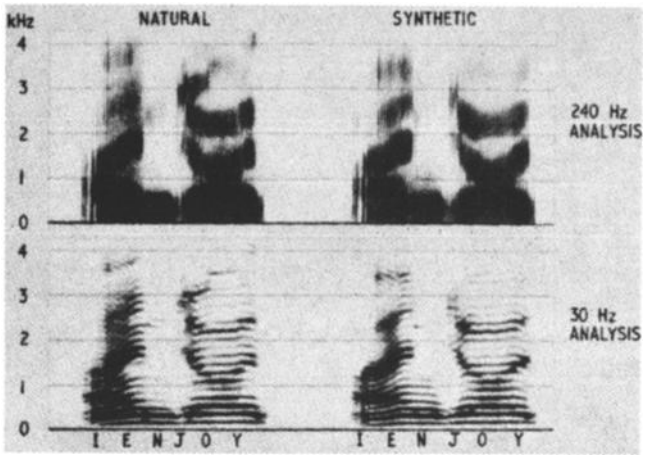


FIG. 9. Comparison of broadband and narrow-band sound spectrograms of a natural utterance and a synthetic imitation produced by Holmes (1973).

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-

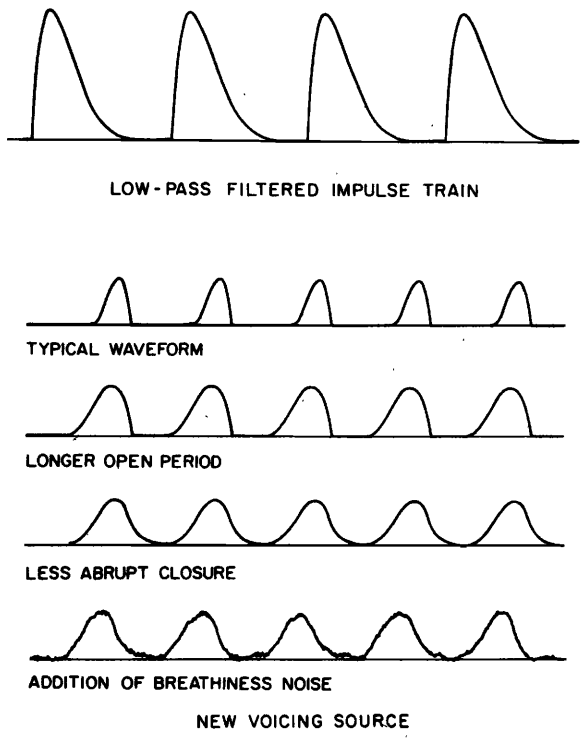


FIG. 10. Comparison of voicing waveforms consisting of a filtered impulse (top) and several more natural waveforms produced by varying the open quotient, spectral tilt, and breathiness in the Klattalk model.

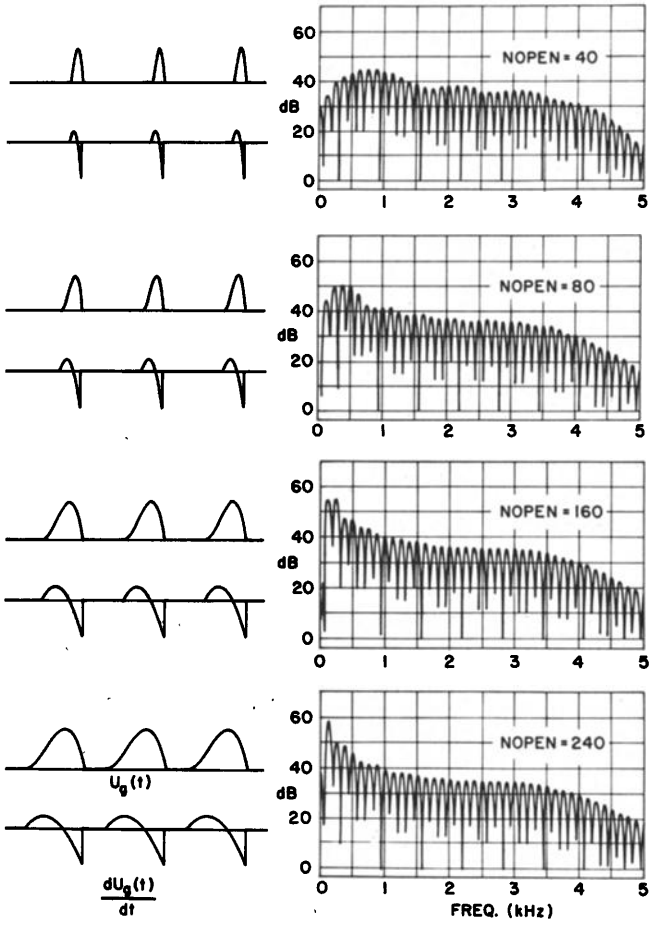


FIG. 11. Selected magnitude spectra of the output of the voicing source model of Klattalk as the duration of the open portion of the glottal cycle is varied from 1 ms (NOPEN = 40) to 6 ms (NOPEN = 240). Note the significant change in relative energy content at low frequencies; the amplitude of the first 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.