CURRENT TRENDS IN LINGUISTICS
Edited by Thomas A. Sebeok
1974. Mouton, The Hague
pp. 2451-2487

SPEECH SYNTHESIS FOR PHONETIC AND PHONOLOGICAL MODELS

IGNATIUS G. MATTINGLY

1. INTRODUCTION

The linguist today, and more especially the phonologist, is aware of two very active areas of investigation. One of them, generative phonology, is at the very center of his field of vision, and cannot be ignored. The other, experimental phonetics, may seem less directly relevant to his concerns, even though he accepts the truism that phonology rests ultimately on a phonetic basis. As modern experimental phonetics grows more technical, and is increasingly dominated by the psychologist, the physiologist, the electronics engineer and the speech scientist, the phonologist is more likely than ever to be put off and to yield to the temptation to do phonology on the basis of phonetic folklore.

In this situation, it is fortunate that there is a group of investigators whose efforts tend to bridge this gap: those engaged in synthesis of speech by rule. By synthesis by rule, we mean the automatic production of audible synthetic speech from a symbolic transcription, by a process which models phonetic and phonological rules in some non-trivial way. With the help of modern electronic technology, it is now perfectly possible to type a transcription on a computer typewriter and immediately hear the corresponding utterance. But the investigator who undertakes synthesis by rule not only makes use of advanced technical facilities, he also attempts of necessity to integrate and generalize into a system those findings of experimental phonetics which are relevant to phonology. On this account, at least, synthesis by rule should intrigue the phonologist.

We believe that there are other reasons as well why the phonologist should be interested, and we shall try to make them clear below. We shall also say a little about the techniques of speech synthesis, give some account of the development of synthesis by rule, describe a number of current approaches to the task, and finally, suggest some possible directions which this work may take in the future.

Speech synthesized by rule is not the only kind of synthetic speech.[1] There are several others which should be mentioned, if only because investigations motivated by these uses have led to technical advances of general benefit. Thus, much

[1] For general discussions of speech synthesis, see Wheatstone 1837; Dudley and Tarnoczy 1950; Fant 1958; Cooper 1962; and Flanagan 1965 : 167-191.

of the speech synthesis research of the past thirty years has been prompted by interest in vocoding (i.e. voice coding). The channel capacity (equivalently, the bandwidth in the radio spectrum) required for transmission of speech is many times greater than it ought to be, considering the amount of information, in Shannon's sense, which is carried by the speech signal. Since the channel capacity available for radio and cable communications is limited, many schemes have been devised to 'compress' speech by analyzing the speech wave and transmitting only the information needed to synthesize an intelligible version at the receiving end. For example, in Dudley's (1939) original Vocoder, built at Bell Telephone Laboratories, the spectrum of telephone speech (250–3000 Hz) is analyzed by a bank of 10 filters. The smoothed, rectified output of each filter represents the energy in a certain part of the spectrum as a function of time. Another circuit tracks F_0, the fundamental frequency (for voiceless excitation, the output of this circuit is zero). The vocoder transmits the outputs of the F_0 tracker and of the filters. Since these functions vary relatively slowly, the channel capacity needed for all 11 functions is far less than the unprocessed speech signal would require. To synthesize the speech, the frequency of a buzz source is varied according to the F_0 function (a hiss source is used when this function has zero value). The buzz or hiss excites each of a set of filters matching those used in the analysis, and the amplitude of the output from each synthesizing filter is determined by the function for the corresponding analyzing filter. Summing the outputs of the synthesis filters yields an intelligible version of the original speech.

A second type of vocoder is the formant vocoder (Munson and Montgomery 1950). In a formant vocoder, the analyzer tracks the excitation state, F_0, and the frequencies and amplitudes of the lowest three formants of the original speech, and transmits these functions; in the synthesizer, resonant circuits representing the three formants are appropriately excited, and the transmitted functions also determine the frequency and the amplitude for each resonator. The saving in channel capacity is greater than for a filter-bank vocoder, but correct analysis is much more difficult. Both filter-bank and formant synthesizers have proved to be of value for phonetic and phonological research as well as for communications.

Besides vocoding, there are certain other possible applications for synthetic speech. If it is necessary for a machine to communicate with its user — a computer operator or a student undergoing computer-assisted instruction — and heavy demands are already being made on his visual attention, spoken messages may be the solution. But if fast random access to a large inventory of messages is required, storage of natural speech becomes cumbersome, for speech makes the same exorbitant demands on storage capacity as it does on channel capacity (Atkinson and Wilson 1968). Synthetic speech, if it could be stored in some kind of minimal representation, would be an attractive alternative. Still another application is a reading-machine for the blind. In such a device, printed text must be converted to spoken output with the aid of a dictionary in which written and spoken elements

are matched. If these elements are naturally spoken words, the output rate cannot exceed ordinary speaking rates without distortion, and (on account of the difficulty of abutting the words closely without losing intelligibility) will actually be considerably slower. Yet the blind user would be happy with an output several times faster than natural speech. A potential solution is suggested by the fact that speech can be synthesized at two or three times normal speed without much loss of intelligibility (Cooper et al. 1969).

In addition to these practical applications, synthetic speech is used for psychological research into the nature of speech perception itself. Synthetic stimuli can be produced which are simple and closely controlled, and which in some cases could not have been produced by a human speaker at all. Such stimuli have been used to study categorical and continuous perception of speech sounds (Liberman et al. 1957), differences between perception of speech and non-speech signals (Liberman et al. 1961) and hemispheric localization of speech processing (Shankweiler and Studdert-Kennedy 1967). These investigations, apart from their intrinsic interest, are an essential preliminary to synthesis by rule.

Another research application of synthesis is the close imitation of natural speech utterances. It is of some interest to know just how faithfully a particular synthesizer can simulate natural speech, without any assumptions being made about the structure of speech or language beyond those built into the synthesizer. Such investigations explore the limitations of the synthesizer. If the best imitations which could be achieved in this way were indeed quite poor, this fact would discourage *any* endeavor making use of synthetic speech. Fortunately, at least one investigator, Holmes (1961; see also Holmes et al. 1964) using a formant synthesizer, has been able to synthesize sentences which are extremely natural and virtually impossible to distinguish from their originals. This is good evidence that progress in other applications of speech synthesis is at any rate not limited by the quality of the synthesizer.

2. HISTORICAL DEVELOPMENT OF SYNTHESIS BY RULE TECHNIQUES

The applications we have just summarized are quite recent. The traditional motivation for research in speech synthesis has been simply to explain how man used his vocal tract to produce connected speech. In a broad sense, such research is synthesis by rule, though it was a long time before the notion of a rule became obvious, and the importance of an explicit formulation of the rules was recognized.

The idea of an artificial speaker is very old, an aspect of man's long-standing fascination with humanoid automata. Gerbert (d. 1003), Albertus Magnus (1198–1280) and Roger Bacon (1214–1294) are all said to have built speaking heads (Wheatstone 1837). However, historically attested speech synthesis begins with Wolfgang von Kempelen (1734–1804), who published an account of his twenty years of research in 1791 (see also Dudley and Tarnoczy 1950). Von Kempelen's

synthesizer was a windbox driven by a bellows. One output of the windbox led to a reed, to simulate vocal-cord excitation; the reed was followed by a short neck and a bell-shaped, rubber mouth. Deforming the mouth changed the quality of the sound made by the reed. Projecting from the neck were two tubes which could be opened to make nasal sounds. Other outputs of the box included special passages for the fricatives [s] and [š]. There was also a lever to modulate the vibration of the reed for trilled [r], and an auxiliary bellows for aspiration of voiceless stops. A human operator played the synthesizer like a musical instrument; he pumped the bellows with his right arm, operated the various levers and tubes with his right hand, and manipulated the rubber mouth with his left hand.

Von Kempelen claims to have synthesized a number of short utterances in various languages ('Leopoldus Secundus', 'vous êtes mon ami'). In 1923, Paget operated a copy of the synthesizer built by Wheatstone (1837) and was able to produce a few isolated words (Paget 1930:19). But whatever the quality of the synthesis, one cannot fail to be impressed by the insights into the nature of speech production reflected in the design of the synthesizer and manifest in von Kempelen's monograph. He understood the basic relationship between the larynx and the supraglottal cavities, and realized the special problems posed by nasals and fricatives. He understood also the importance of what Fant et al. (1963) were later to call 'synthesis strategy': a set of techniques for producing the various classes of sounds which exploits the possibilities of a particular synthesizer and minimizes its limitations (thus von Kempelen made an [f] by closing all regular outlets from the windbox and building up enough pressure inside to force air through the leaks in the box!). The obvious limitations of his work were the need to use a mechanical system, the acoustic properties of which were neither easily predictable nor readily alterable; and the use of a human operator for dynamic control, with the consequence that the 'rules' were not explicit, but were rather part of the operator's art. Interestingly enough, the Abbé Mical, a contemporary of von Kempelen's, is supposed to have built a synthesizer controlled by a pinned cylinder, such as is used in a music box. Wheatstone (1837:40–1) considered and dismissed the possibility of fitting his copy of von Kempelen's synthesizer with a control device of this sort:

> It would be a very easy matter to add either a keyboard or a pinned cylinder to De Kempelen's instrument, so as to make the syllables which it uttered follow, each with their proper accentuations and rests; but unless the articulations were themselves more perfect, it would not be worth the trouble and expense.

On the other hand, without a well-specified input, such as a pattern of pins on a cylinder, how can the performance of the synthesizer be systematically studied and improved?

During the century after von Kempelen and Mical, other manually operated, mechanical speech synthesizers were developed. Besides Wheatstone's copy of von Kempelen's synthesizer there was, for instance, Faber's Euphonia (Gariel 1879),

which according to Dudley and Tarnoczy (1950) had variable pitch and sang 'God Save the Queen'.

But the next real step forward was Dudley's Voder (Dudley et al. 1939), an offshoot of his Vocoder, described earlier. The ten filters of the Vocoder synthesizer were widened to cover the range 0–7500 Hz and a set of manual controls was supplied. The output amplitudes of the filters were controlled from a keyboard; a wrist bar selected buzz or hiss excitation and a foot-pedal controlled F_0. There were also special keys to generate automatically the sequence of closure and release required for stops. A year or more of training was required before an operator could produce intelligible speech, and each utterance had to be carefully rehearsed. The Voder was demonstrated successfully at the 1939 New York World's Fair and the 1940 San Francisco World's Fair.

There were two important differences between Dudley's approach to speech synthesis and von Kempelen's. First, Dudley's Voder was an electrical rather than a mechanical simulation, so that the acoustic properties of synthesizer components were reasonably predictable and design changes could be readily made. Second, the Voder simulated acoustic properties of speech whereas von Kempelen's simulated articulatory properties as well. Dudley's model made it easier to improve the rendition of particular speech sounds, but made a weaker claim about the nature of speech. However, Dudley's system had one important feature in common with von Kempelen's: a human operator was used, and the rules for synthesis were part of his skill.

The human operator disappeared from speech synthesis as an indirect result of R. K. Potter's invention of the sound spectrograph during World War II (Koenig et al. 1946). After the War, the spectrograph opened the way to extensive research in acoustic phonetics because it made it easy to observe the correspondence between speech sounds and events in the acoustic spectrum, notably formant movements. The spectrograph also suggested a new way of synthesizing speech: 'playing back' a spectrogram. Potter himself built a playback synthesizer (Young 1948); Cooper (1950) developed a research version, the Pattern Playback, which is still in use at Haskins Laboratories. In the Pattern Playback, an optical representation of an excitation spectrum with 50 harmonics and F_0 at 120 Hz is shaped by a spectrographic pattern painted on a moving, transparent acetate belt; and this optical representation is then converted to an acoustic signal. Thus the synthesis of an utterance is not a transient performance but is controlled by a pre-planned pattern, and can be repeated. Moreover, the close correspondence between the output of the analyzing tool (the spectrograph) and the input to the synthesizing tool (the Playback) is convenient experimentally and of great value conceptually.

The Haskins investigators used the Playback to study the psychology of speech perception and to accumulate a body of knowledge about the 'speech cues' (Liberman et al. 1967). Experienced users of the Playback, for example the late Pierre Delattre, could readily paint intelligible utterances; like the operators of von Kem-

pelen's synthesizer or the Voder, they had internalized a set of rules. Frances Ingemann, however, used this body of knowledge to draw up a formal set of instructions for painting spectrograms (Ingemann 1957; Liberman et al. 1959). Her instructions are subdivided into rules for manner class, place of articulation and the voiced-voiceless distinction (since the rules are intended for a monotone synthesizer, there are no F_0 rules). The manner-class rules specify steady-state durations, the durations of the associated formant transitions, and the formant amplitudes appropriate for each class. The place rules specify the steady-state frequencies of formants, fricative noise and stop bursts, and the transition endpoints for stops, nasals, fricatives and affricates at each point of articulation. The voicing rules specify the burst durations and closure voicing for stops and the friction duration and intensity for fricatives. Because they are organized along the phonetic dimensions of manner, voice and place, the rules make the most of the uniformities and symmetries which emerged from research on the speech cues. But this kind of organization could not be carried through consistently: separate rules were needed for the steady states of each vowel, and 'position modifiers' — changes to the basic rules — were required in some contexts. These modifiers reflect basic limitations of synthesis by rule at the acoustic level.

Using these rules, the utterance, 'I painted this by rule without looking at a spectrogram and without correcting by ear. Can you understand it?' was synthesized on the Playback. Painting an utterance of any length to precise specifications, however, proved to be a laborious procedure, and not many other such utterances were actually synthesized. Nevertheless, for the first time, a set of rules had been stated explicitly enough so that anyone willing to take the trouble could paint a spectrogram by rule and synthesize the corresponding utterance, any spectrogram purporting to follow the rules could be checked, and utterances representing different versions of a rule could be directly compared. The concept of speech synthesis by rule, which had been the implied purpose of earlier investigators, now became a clearly understood research objective.

Meanwhile, the resonance synthesizer had been developed. This type of synthesizer derives from the formant vocoder just as the Voder derives from the filter-bank vocoder. Resonant circuits represent the first few formants and these circuits are excited by a hiss source or a variable-frequency buzz source. The state of the synthesizer can thus be specified by assigning values to a relatively small number of parameters which correspond to significant dimensions of natural speech and are readily observable in acoustic records: F_0, the formant frequencies F_1, F_2, F_3, and so on. A much stronger claim is implicit in a parametric synthesizer than in a non-parametric synthesizer like the Voder or the Playback. The information of interest in speech is regarded not just as band-limited, but as entirely a function of a very few physical variables.[2]

[2] For an interesting discussion of parametric vs. non-parametric synthesis, see Ladefoged 1964.

The resonant circuits of the synthesizer can be arranged either in parallel, the outputs of the various circuits being summed to produce the final output, or in series, the output of each resonant circuit being fed into the next. The series synthesizer is a closer approximation to the acoustical behavior of the vocal tract, and incorporates in its design Fant's (1956) observation that if certain assumptions are made about the glottal source spectrum and the formant bandwidths, the relative amplitudes of vowel formants can be predicted from their frequencies. Thus a series synthesizer requires fewer parameters and can be expected to produce more natural vowels. On the other hand, parallel synthesizers are far more flexible, and simplify synthesis strategy for sounds with complex spectra, like voiced fricatives. The relative merits of parallel and series synthesizers are far best summed up by Flanagan (1957). Lawrence's (1953) PAT was the first example of a parallel resonance synthesizer; Fant's (1958) OVE II, the first full-scale experimental series synthesizer. Highly reliable resonance synthesizers of both types are now available.[3]

A parametric control scheme should have made synthesis by rule simpler, since the parameters to be specified are precisely the dimensions of speech in terms of which it is convenient to state acoustic rules. Ingemann (1960), in fact, reformulated her rules for use with the Edinburgh series version of PAT (Anthony and Lawrence 1962). But in order to control a resonance synthesizer, some means of changing the parameter values dynamically is required. At first this was accomplished with a function generator: for example, parameter functions for the Edinburgh PAT were represented in conductive ink on parallel tracks of a moving plastic belt (Fourcin 1960). But *applying* the rules (as distinct from *stating* them) was if anything more troublesome with a function generator than with the Pattern Playback. Fortunately, digital computers now began to become available for phonetic research. Kelly and Gerstman (1961) demonstrated that the computer not only could apply a set of rules (i.e., calculate the parameter values) quickly and accurately but also could be used to simulate the synthesizer itself.[4] Other investigators showed that if an actual, rather than a simulated synthesizer is used, the computer could also play the role of function generator.[5]

[3] Other parallel resonance synthesizers are described in Borst 1956, Holmes et al. 1964, Mattingly 1968b and Glace 1968; other series synthesizers are described in Coker 1965, Tomlinson 1965, Liljencrants 1968, Kacprowski and Mikiel 1968, Kato et al. 1968, Dixon and Maxey 1970, Shoup, pers. comm.

[4] The advantage of a simulation is that it can be completely reliable and accurate, and the design of the synthesizer can be readily modified; the disadvantage is that an extremely powerful computer is required and such computers are too expensive to permit extended real-time operation. Recent simulations of resonance synthesizers (all series) include those described in Flanagan et al. 1962, Rao and Thosar 1967, Rabiner 1968, Saito and Hashimoto 1968.

[5] On-line transmission of stored parameter values can be performed by a laboratory computer at a cost low enough to permit the investigator to experiment at length; it is easy to program other convenient facilities such as routines for editing or displaying the stored parameter values. Schemes of this sort include those of Tomlinson (1965), Denes (1965), Coker and

The Kelly and Gerstman program was quite simple. For each speech sound, initial and final transition durations, steady-state durations, and steady-state values for each parameter were stored. During the steady state of a sound, the stored values were used; during the final-initial transition period, parameter values changed smoothly from preceding to following steady state. It would be easy to criticize this scheme: the framework within which the rules are stated is extremely crude, and a good deal of ad hoc modification was required to make the synthetic speech even reasonably intelligible. But Kelly and Gerstman had clearly demonstrated that a computer could be used to apply phonetic rules — as great an advance over application of the rules by drawing patterns or functions by hand as were the latter over direct operation of the synthesizer by a human being. It was now possible to test and correct rules by producing substantial quantities of synthetic speech automatically and consistently.

Resonance synthesizers, as well as the Playback and the Voder, are 'terminal analog' synthesizers: they simulate the acoustic output of the vocal tract but not the activity of the vocal tract itself. However, concurrently with resonance synthesizers, vocal tract analog synthesizers were being developed. With one interesting exception, the 'true' (i.e. mechanical) model of Ladefoged and Anthony (Anthony 1964), these synthesizers are electrical simulations. The supraglottal vocal tract is considered as segmented into a series of short tubes, each with variable cross-sectional area. The acoustic properties of each tube in such a series can be simulated by the electrical properties of a transmission line. The acoustical effect of a change in cross-sectional area is equivalent to a change in the characteristic impedance of the corresponding transmission-line segment. The nasal cavity is usually represented as a branch with a few fixed sections and variable coupling to the main line. Thus, the spectrum of the output of the synthesizer depends on the momentary cross-sectional area function and the amount of nasal coupling.[6]

Like a resonance synthesizer, a vocal-tract analog could be simulated on a computer, and Kelly and Lochbaum (1962), used such a simulation for synthesis by rule. The approach was very much the same as the one used by Kelly and Gerstman, except that the parameters were the areas of the cross-sections of the segments of the tract instead of formant frequencies. The results were less successful than Kelly's terminal-analog synthesis had been; a fact of some interest.

Cummiskey (1965), Scott et al. (1966), Mattingly (1968b). Off-line control schemes, in which the computer produces a record, such as a paper tape, which is then used to control a function generator, are also practical, though less convenient (Holmes et al. 1964; Iles 1969).

[6] The first electrical vocal tract analogs were static, like those of Dunn (1950), Stevens et al. (1953), Fant (1960). Rosen (1958) built a dynamic vocal tract (DAVO), which Dennis (1963) later attempted to control by computer. Dennis et al. (1964), Hiki et al. (1968) and Baxter and Strong (1969) have also described hardware vocal-tract analogs. Kelly and Lochbaum (1962) made the first computer simulation; later digital computer simulations have been made, e.g. by Nakata and Mitsuoka (1965), Matsui (1968) and Mermelstein (in press). Honda et al. (1968) have made an analog computer simulation.

By the early 60s, then, there was no doubt that speech could be synthesized by rule by either terminal-analog or vocal-tract analog methods. Reliable synthesizers and convenient methods of controlling them had been developed. Even more important, the value of explicitly formulated rules had become obvious.

3. JUSTIFICATION FOR SYNTHESIS BY RULE

Since speech has been successfully synthesized by rule, it might seem that the basic objective of von Kempelen and his successors has been attained; it therefore becomes important to state clearly the reasons for continuing the research. One obvious reason is that we still have much to learn about the physical aspects of speech production: how the various articulators move, how their movements are timed; how the controlling musculature operates to produce the sounds of speech. Synthesis by rule is a way of testing our understanding of the physical apparatus, and this is the primary motivation for much of the activity in the field today. But this argument may not seem very persuasive to the linguist, who is concerned with speech as a psychological fact rather than a physical one. But we think that synthesis by rule offers other possibilities of substantial theoretical importance for the linguist, possibilities which have barely begun to be explored. To justify this point of view, however, requires brief reference to some basic questions of linguistics.

We believe, following Chomsky and Halle (Chomsky 1965, 1968; Chomsky and Halle 1968) and other generative grammarians, that the grammar of a language can be represented by a set of rules. A speaker-hearer who is competent in a language has learned these rules, and uses them to determine the grammatical structure of his utterances or those of another speaker, since the rules 'generate' an utterance if and only if grammatical structure can be assigned to it. Competence does not fully determine performance: the speaker's actual utterances may frequently be ungrammatical, and the listener may guess a speaker's intent without consistent reference to the rules.

A subset of the rules for any language are phonological: they convert a string of morphemes, already arranged in some order by syntactic rules, into a phonetic representation. In the familiar generative model (Chomsky and Halle 1968), the morphemes are lexically represented as distinctive-feature matrices, each column of which is a phonological segment. All phonologically redundant feature specification is omitted, and each specified feature has one of two values. The phonological rules complete the matrices, alter the feature specification in certain contexts, delete and insert segments, and assign a range of numerical values to the features. The output of the phonological rules, then, is a matrix of phonetic segments for which each of the features is numerically specified.

Besides these acquired rules, we suppose the speaker of the language to have

a certain inborn linguistic capacity: 'the innate organization that determines what counts as linguistic experience and what knowledge of language arises on the basis of this experience' (Chomsky 1968:24). This capacity is what makes it possible for him to learn the rules and to use them in making grammatical judgments; it is reflected in some universal and quite severe constraints on the form and content of grammatical rules. In the case of phonology, not only are the forms of the input and output highly determined, but the set of phonetic features by which the output of the phonology may be represented is the same for all languages; moreover, the language-specific set of 'classificatory' features which are used to represent the lexical items at the input to the phonology are related in a significant though complex way to a subset of universal phonetic features with the same names. Phonological rules are constructed out of phonetic raw material.

Conventionally, the linguist's concern ends with the phonetic-feature representation which is the output of the phonology. But our account of the speaker-hearer's inborn capacity is incomplete, for we have said nothing about his knowledge (quite unconscious, but no less psychologically real) of the relationship between the phonetic-feature representation and the acoustic signal. The speaker-hearer can produce an acoustic representation of an utterance, given the feature representation (speech production), and he can apparently recover the feature representation for an utterance produced by someone else (speech perception). Neither process is trivial; to be able to produce and recognize speech, he must possess a definition of each feature in sufficient detail to enable him to assign a possible value to it, given the acoustic signal. He must therefore have sufficient information about the anatomy, physiology and acoustics of the vocal tract to permit such a definition; and he must also understand how the apparently discrete representation of an utterance, at the output of the phonology, as a series of phonetic segments, is translated into a continuous representation in the acoustic signal. Perhaps it would not be inappropriate to picture the speaker-hearer's knowledge as consisting of a dynamic neural simulation of the vocal tract, the state of which is determined by the values of the features, and which guides his production and perception of speech (Mattingly and Liberman 1969). If we assume that both the features and the general structure of the human vocal tract are universal, it seems highly likely a priori that this knowledge of the speaker-hearer's is inborn; moreover, some experimental evidence for such a view has recently appeared (Moffitt 1969; Eimas et al. 1971).

Just as we characterize phonological and syntactic capacity by assigning to them formal and substantive properties, in the form of conventions and features, so, to the extent that we can describe this simulated vocal tract, we characterize what may be called 'phonetic capacity'. It is to phonetic capacity which Chomsky and Halle (1968:294–295) allude when they observe:

> The total set of features is identical with the set of phonetic properties that can in principle be controlled in speech: they represent the phonetic capabilities of man and, we should assume, are therefore the same for all languages.

We view the development of an adequate account of phonetic capacity as the chief goal of experimental phonetics. There appear to be two important tasks. First, it is necessary to determine the membership of the set of universal phonetic features, since these are the basis of phonological capacity and the elements of phonological competence. Moreover, we want to understand the role of the various stages in the speech chain in the psychological definition of each feature. These stages include (at least), the activation of the muscles of the vocal tract by neuromotor commands, the gestures made in response to these commands by the various articulators, the resulting dynamic changes, not only in the shape of the vocal tract but also in air pressure and airflow at different points in the tract, and finally the cues in the acoustic output. It may well be that not all these stages are pertinent to phonetic capacity; on the other hand, other stages, as yet poorly understood, may be involved.

The second task is to characterize psychologically the translation from the discrete, essentially timeless phonetic level to the continuous, time-bound activity characteristic of lower levels. If at any of these levels speech could be consistently separated into stretches corresponding to phonetic segments, the problem would be fairly simple; but we know that this cannot be done. At any level in speech which we can observe, there are no true boundaries corresponding to any phonetic unit shorter than the breath-group, though in the acoustic signal there are many apparent boundaries reflecting articulatory events, e.g. stop closures and releases, spectral discontinuities in liquid and nasal sounds, onset and offset of voicing and the like. Yet the psychological reality of phonetic segments can hardly be doubted; and at any rate, without them, phonology would collapse.

The feature-specified, dynamic vocal tract model by which we would represent phonetic capacity is on just the same level of theoretical explanation as the phonological and syntactic models of linguistics. It does not have, as yet, any but the most general sort of neurophysiological basis; it does not account in itself for productive or perceptual performance; in particular, it does not conflict with an analysis-by-synthesis account of speech perception. It is simply a way of stating some properties which the neural mechanisms for speech must incorporate to account for the observed behavior of speaker-hearers.

By virtue of his phonetic capacity the speaker-hearer acquires certain phonetic skills, just as he acquires the phonological rules of his language by virtue of his phonological capacity. He must 'calibrate' his perceptions to allow for the idiosyncrasies of the vocal tracts of the other speakers to whom he listens, even before he understands his native language. (If there were no other reason for postulating phonetic capacity, we would want to do so to account for the fact that an infant learns to interpret the output of the vocal tract of each of the individuals around him, though these vocal tracts differ radically from one another and from his own in size and shape; and he does so with sufficient accuracy to permit the collection of the 'primary linguistic data' (Chomsky 1965:25) essential for language ac-

quisition). Moreover, if he himself is to produce acceptable versions of the speech sounds he perceives, he also has to learn the idiosyncrasies of his own vocal tract. He must learn to control certain stylistic factors — speaking rate, attitudinal intonation and so on, both in production and perception. Finally, he may need to learn certain global phonetic properties of his language, e.g. its 'articulation basis' (Heffner 1950:98–9).

In brief, we distinguish four distinct components underlying speech perception and production: 1) inborn phonological capability, 2) acquired phonological competence in one's language, 3) inborn phonetic capacity, 4) acquired phonetic skill.[7] For the study of these various components, speech synthesis by rule has certain impressive advantages.

First, we can hope to gain real understanding of the component of interest to us only by attempting a highly formal account; yet any nontrivial formal account will doubtless be quite complex: this is already apparent for phonological and phonetic capacities and particularly so for phonological competence — as a glance at the summary of rules in Chapter 5 of *The sound pattern of English* (Chomsky and Halle 1968) will confirm — and must certainly prove true for phonetic skill as well. Much can be done to reduce apparent complexity by suitable notation. But, as in many other fields which make use of highly formal systems, checking the consistency of the formalization is most easily done by computer simulation. Linguists are in fact turning increasingly to computer simulation to check the operation of syntactic and phonological rules (Fromkin and Rice 1970).

Second, various dependencies exist among the components. An account of the phonetic skill of a particular speaker must begin with some assumptions about his phonological competence in his language and his phonetic capacity. Only in terms of the former can idiolectal variations be defined; only in terms of the latter can speaking rate be discussed. Similarly, the rules by which we try to characterize phonological competence must be stated in a form determined by phonological capacity. Phonological capacity, finally, depends on the choice of a set of features, the interpretation of which is a matter of phonetic capacity. Given this kind of dependency, it seems extremely risky to try to form hypotheses about the nature of one component without being quite specific as to the assumptions being made about the others on which it depends. Yet this is an ever-present temptation. Speaker variation, for example, is investigated without specification of precise phonetic and phonological models. Structural linguists rightly incurred the censure of generative phonologists because they formulated their phonemic inventories without proper concern for phonological capacity; generative phonologists, in turn,

[7] Tatham (1969a) has recently used the term 'phonetic competence' to mean approximately what we mean by 'phonetic capacity'; otherwise we might have used the former term rather than the assymetrical 'phonetic skill'. Tatham's paper (see also Tatham 1969b) contains some cogent arguments not only for the existence of phonetic capacity but also for its importance in the formulation of phonological rules in a natural way.

might be criticized because the set of phonetic features, on which their much more principled account of phonological capacity depends, as yet lacks a fully satisfactory and explicit basis in phonetic capacity (Abramson and Lisker 1970). Obviously, it is very desirable to state clearly, when a certain component is being investigated, how this component is assumed to depend on other components.

Third, the ultimate check of a hypothesis concerning any or all of the components is of course the intuition of the native speaker (Chomsky 1965:21). However, the only reliable way to consult his intuition is to present him with speech which we have made sure conforms to our current phonetic or phonological hypothesis and find out whether he considers it well-formed. To do this, however, we need carefully controlled speech stimuli (Lisker et al. 1962; Mattingly 1971).

Synthesis by rule is a technique which seems to meet these requirements. With the computer we can simulate our phonological and phonetic formulations rigorously; errors of form and logic come to light all too quickly. We are compelled to be explicit about the assumptions we make about other components; if they are simplistic or inadequate we will not be allowed to forget the fact. And we can check the native speaker's intuition directly by producing controlled synthetic speech.

Let us briefly consider what an ideal speech synthesis by rule system would be like. It would, in the first place, simulate all the components we have just discussed. Phonetic capacity would be represented by a synthesizer and computer programs controlling it which are capable of generating just those sounds which can be distinguished in production and perception by the speaker-hearer; phonological competence, by the rules of some language, stated in a form which would be an acceptable input to the system; phonological capacity, by a part of the computer program itself, which would impose severe limitations on the form or substance of the rules; and phonetic skill, by an additional set of rules specific to some particular speaker. The combined effect of all components should be such as to restrict the possible utterances to just those which are well-formed speech in a particular language (assuming appropriate syntactic and semantic constraints) from one particular speaker to another.

For each component, moreover, we would want to include all those aspects, and only those, which are relevant to the capacity and competence underlying the production and perception of speech. Suppose, for instance (contrary to our present expectations) that, from a psychological standpoint, speech production proved to be only a matter of transmitting certain cues definable in acoustic terms and invariantly related to phonetic features, and that speech perception consisted simply in detecting these cues. Our 'neural vocal tract simulation' could then be just a terminal analog synthesizer. There would then be no reason for including neuromotor commands, gestures or shape change in a parsimonious synthesis by rule system, because these matters would be irrelevant to phonetic capacity. They might continue to be of great interest from the standpoint of the physiologist and acoustician interested in speech, but would have no claim on the linguist's attention.

Our ideal system is not concerned with performance as such. Even though our model is dynamic and the output is audible, the process of synthesis is a derivation according to rules, not a life-like imitation of a speaker's actual speech behavior. The output is acceptable to the hearer because it follows the rules, not just because, on the one hand, it is intelligible, despite errors and deviations, or on the other, because it is highly natural-sounding — though one might expect that the output of an ideal system would be natural-sounding, if not physically naturalistic. Here our emphasis differs somewhat from that of Ladefoged (1967) and Kim (1966) who share our conviction that it is important to do synthesis by rule, but for whom linguistic and phonetic theory 'must lead to the specification of actual utterances by individual speakers of each language; this is physical phonetics' (Ladefoged 1967:58). From our point of view it is not physical realism but psychological acceptability which is the proper evidence for correctness at the phonological and phonetic levels, just as it is on the syntactic level.

In the preceding discussion, we have deliberately generalized the concept of 'synthesis by rule' to embrace phonology and phonetics. It would be possible to generalize still further, to include syntax and semantics in a synthesis by rule system. But while computer simulations of syntactic and semantic rules are certainly desirable, the motivation for coupling them to a phonological and phonetic synthesis by rule system is less compelling, primarily because a set of syntactic rules can in practice be evaluated more or less independently of the associated phonology and phonetics.

4. CURRENT WORK IN SYNTHESIS BY RULE

We turn now to an assessment of the progress which has been made toward the ideal which has just been sketched. The first thing to be said is that most of the activity and most of the progress so far falls under the heading of phonetic capacity. Since the other components all depend, directly or indirectly, on phonetic capacity, this is just as it should be. Moreover, since we want to assess the role of the different stages of the speech chain in phonetic capacity, it is good that, in the present state of our knowledge, the research has been pluralistic: different types of systems have been developed in which the contribution of different stages has been emphasized. This has been difficult to do because appropriate data on which to base investigations at stages before the acoustic stage are hard to collect. At present, most of the work has been at the acoustic stage; the relationship between shape and acoustic output is quite well understood and several synthesis-by-rule systems operating on vocal-tract shape have been developed; systems which represent the movements of the actual articulators are beginning to show results; and some work has been done at the neuromotor command stage.[8]

[8] There is, of course, another way to synthesize speech by rule, and that is to compile an utterance from an inventory of shorter segments, themselves either natural or synthetic. Such

4.1 *Acoustic-level Systems*

We have already mentioned some systems in which the phonetic level is mapped directly on to the acoustic level, including one, that of Kelly and Gerstman (1961), which, like other more recent systems of this kind, is parametric. In these systems a target spectrum for each phone[a] is specified by a set of stored parameter values. Given a phonetic transcription of an utterance, the synthesis program calculates the momentary changes of value for each parameter from target to target as a function of time. (Notice that this is an extremely natural way to treat the problem of translating from the discrete to the continuous domain.)

The most important differences among the various systems have to do with the procedures for this calculation, and in particular, the procedure for calculating formant motion, since intelligibility depends crucially on the choice of targets toward which the formants move, and the timing of their movements. In the Kelly-Gerstman system, it will be recalled, an initial transition duration, a final transition duration and a steady-state duration are stored for each phone. The duration of a transition between two adjacent phones is the sum of the final transition duration of the first phone and the initial transition duration of the next. During the steady-state period, formants remain at their target values; during the transition period, they move from one set of target values to the next, following a convex path from consonant to vowel, a concave path from vowel to consonant, and a linear path otherwise.

In the system of Holmes et al. (1964), a 'rank' is stored for each phone, corresponding to its manner class. Manner classes having characteristic transitions (e.g. stop consonants) rank high; manner classes for which the character of the transition is characterized by the adjacent phone rank low. The character of the transition between adjacent phones is determined according to the ranking phone. Each transition is calculated by linear interpolation between a target value for the first phone and a boundary value, and between the boundary value and the target for the second phone. The durations of the two parts of the transition are stored for the ranking phone. The boundary value is equal to $C_R + W_R (F_A)$, where C_R is a constant and W_R a weighting factor for this formant stored for the ranking

approaches may have practical value, but from a theoretical standpoint they merely serve to remind us that there is no simple correspondence between phones and segments of the acoustic signal. See the discussion in Liberman et al. 1959. Systems in which speech is compiled from natural segments have been described in Harris 1953, Peterson et al. 1958, Cooper et al. 1969. Systems using synthetic segments are described in Estes et al. 1964, Dixon and Maxey 1968 and Cooper et al. 1969.

[a] Workers in synthesis by rule (including the author) have been in the habit of referring to the units of their input transcriptions as 'phonemes'. In most cases, these units do not correspond either to the phonemes of structural linguistics or to the phonological segments of generative phonology; they tend to be closer to the level of a broad phonetic transcription. We use the term phone except in the case of systems where a deliberate distinction is attempted between phonological and phonetic levels.

phone, while F_A is the target value of this formant stored for the adjacent phone. Hence, the character of the transition depends mainly on variables stored for the ranking phone. Thus, each phone has within its boundaries an initial transition, influenced by the previous phone, and a final transition, influenced by the following phone. A duration is stored for each phone; if it is greater than the sum of the durations of the initial and final transitions calculated for the phone, the target values are used for the steady state portion. If the duration is less than the sum, and the paths of the calculated transitions fail to intersect, they are replaced by a linear interpolation between the initial and final boundary values. But if the paths do intersect, the values for each transition between the boundary value and the intersection are used, and the others discarded. Thus the formants of shorter vowels do not attain their targets; their frequencies are context-dependent, as in natural speech (Shearme and Holmes 1962; Lindblom 1963).

Denes (1970) uses a similar scheme, the boundary values being dependent on the target values and on a weight assigned to each phone. Our own system (Mattingly 1968a, b) also uses a scheme like that of Holmes et al., except that interpolation is done according to a simple non-linear equation which assures that formants curve sharply near boundaries. The formant transitions in Rabiner's (1967) system, the most serious attempt to simulate natural formant motion, are calculated according to a critically damped second degree differential equation. The manner in which a formant moves from its initial position towards the next target depends on a time constant of the equation, which is specified for each formant and each possible pair of adjacent phones. When all formants have arrived within a certain distance of the current target, they start to move toward the following target, unless a delay (permitting closer approximation or attainment of the target) is specified. It is not obvious that schemes for non-linear motion offer any great advantage over linear schemes. While a non-linear rule results in formant movements which are more naturalistic, they do not seem to be necessarily perceptually superior to, or even distinguishable from, linear movements. If the formant moves between appropriate frequencies over an appropriate time-period, the *manner* of its motion does not seem to be too important.

In Rao and Thosar's (1967) system, each phone is characterized by a set of 'attributes', i.e. features of a sort. A phone is either a vowel or a consonant; vowels are front or back; consonants are stops or fricatives; voiced or unvoiced; labial, dental or palatal. Transition patterns depend on these attributes and on the duration and steady-state spectral values stored for each phone. Vowel-vowel transitions are linear from steady state to steady state, and the two temporal variables — total transition time and the fraction of the total within the duration of the earlier vowel — are the same for all pairs of vowels. For consonant-vowel transition, the boundary value for each formant is equal to $F(F_L) + (1-F)F_V$, where F_V is the target frequency of the vowel, F_L is the consonant locus frequency and F is a weighting factor. Transition time and F_1 locus depend on the value

of the stop-fricative attribute; F_2 and F_3 loci and the weighting factor, on the place-of-articulation attribute. Given the boundary value, steady-state values and transition times, transitions are calculated as by Holmes et al.

Rao and Thosar resort to stored values for vowel spectra and vowel durations; Kim (1966), however, proposes that even these matters can be systematically treated. For example, his translation from distinctive feature values to formant frequencies is made by defining the features in terms of 'degrees' of difference from the [ə] frequencies. From the value assigned to one degree, and the [ə] frequencies, the frequencies of other vowels are calculated by means of such rules as 'if High, –2d'. The formant frequency values determined in this way agree well with the data in the literature. However, since the degree values are not predicted on any principled basis, but are arrived at inductively by an averaging procedure applied to this same data, the agreement is hardly surprising and does not represent any interesting advance over stored values.

Several of these systems have been empirically successful in that they have proved capable of consistently producing intelligible speech. They also have enough theoretical plausibility to be used in investigations of other components. One could, for example, use them to test phonological rules proposed for a language (Mattingly 1971). But they are still inadequate because their working assumption is that phonetic capacity can be adequately described at the acoustic level. If this were so, a simple and consistent correspondence would hold between phonetic features and acoustic events. But in fact the correspondence is only partial. On the one hand, certain regularities are observable, which can be exploited in a synthesis-by-rule system, as Liberman et al. (1959) pointed out: F_1 and F_2 transitions and the type of acoustic activity during stop closure provide a basis for a purely acoustic classification of labial, dental and velar voiced stops, voiceless stops and nasals. On the other hand, the cues for a particular feature, regarded simply from an acoustic standpoint, are a rather arbitrary collection of events. There seems no special reason why a fall in F_1, a 60–150 msec. gap, a burst, and a rise of F_1 should all be cues for a stop consonant, and no obvious connection between the locus frequency and the burst frequency of a stop at the same place of articulation. These cues only make sense in articulatory terms. Still, the apparent arbitrariness of the cues should not in itself discourage the formulation of acoustic rules for features. A more serious difficulty is that in many cases features cannot be independently defined at the acoustic level. Thus the voiced-voiceless distinction is cued in one way for stops and in another for fricatives. The frequencies at which noise is found in a fricative do not correspond to the frequencies of either the locus or the burst of a stop at a similar point of articulation. The frequencies of the first and second formants are sufficient to distinguish the non-retroflex vowels, but the range of F_1 variation seems to be influenced by the F_2 value: the vowels are not distributed regularly in F_1/F_2 space. Because of these difficulties most of the acoustic synthesis by rule systems provide only for a regular relationship between

phones and acoustic events; they do not attempt to define a set of acoustic features. The simple system of Rao and Thosar is exceptional in that (like Ingemann's set of Playback rules) it tries to make the most of the regularities which do exist; but the idiosyncratic characteristics of each phone must also be specified by these investigators.

The manner in which these systems translate from the discrete phonetic level to the continuous acoustic level also proves somewhat unsatisfying. The notions 'target' and 'transition' imply that the former characterizes essential aspects of a phone and the latter is a means of connecting one phone smoothly with another. In fact, as is well-known, much of the information at the acoustic level in speech is encoded in the formant transitions, and most of the ingenuity devoted to acoustic rules has had the purpose of providing appropriate transitions for the various form and manner classes. This circumstance does not invalidate the notions 'target' and 'transition' for synthesis by rule in general; it is merely a further indication of the inadequacy of acoustic synthesis by rule.

4.2 *Vocal-tract Shape Systems*

It appears, then, that there are limitations on the adequacy of a synthesis-by-rule system operating only with the acoustic stage. A number of systems have therefore been developed which incorporate earlier stages in the speech chain.

The next earlier stage in the speech chain is vocal tract shape, which, for a given source of excitation, determines the spectrum of the acoustic output (Fant 1960). Since the acoustics of speech production is complex, it seems plausible that rules for synthesis could be more readily and simply stated in terms of dynamic variations in shape. The speech implied by a sequence of shapes can then be heard with a vocal-tract analog synthesizer.

The general strategy used for synthesis by rule with a vocal tract analog, which parallels the strategy used for acoustic systems, has been to specify a target shape for each phone and to interpolate by some rule between targets. In the system of Kelly and Lochbaum (1962), transition times and target shapes, represented as area functions, are stored for each phone. During the transition, the series of area values for each segment of the vocal tract analog (and also values for excitation parameters and nasal coupling) are obtained by linear interpolation between the target values. There are numerous exceptions to this general principle of operation, most of which are attempts to provide for the effects of coarticulation and centralization. Vowels next to nasal consonants are nasalized throughout. Labials do not have a fixed target shape: the lips are constricted or closed for a period, during which the rest of the tract moves from the previous to the following target. An unstressed vowel has zero duration and its target shape is the average of the shape for the corresponding stressed vowel and that for the neutral or [ə]

vocal tract shape. Separate target shapes are provided for velars before front, central and back vowels.

Mermelstein's (in press) system follows a similar plan. Two lists serve as input to this system. The first is a table of the area function values for an inventory of shapes; the second includes a series of target shapes, specified with reference to the first list and corresponding to phones or temporal segments of phones; target values for other parameters; and transition durations. Mermelstein uses linear transitions near sharp constrictions and exponential transitions during periods when the shape of the tract is changing more slowly. He finds that this procedure, which effectively avoids steady states, contributes considerably to the naturalness of the speech.

Nakata and Mitsuoka (1965) use a more elaborate transition procedure based on a conception of Öhman (1967). In the case of vowel-to-vowel transitions over a period t', the momentary area function for a vowel at αt,

$$^{v}A\,(t') = {}^{v}A^{T} + ({}^{v}A^{0} - {}^{v}A^{T})\ W_{E}\,(\alpha t')$$

where $^{v}A^{0}$ is the starting value, $^{v}A^{T}$ is the target value, and $W_{E}\,(\alpha t')$ an asymptotic weighting function equal to 1 at starting and 0 at target. In the case of a consonant between two vowels, the effect of superposition of the consonant is taken as equivalent to the effect over a period θ of the consonant on the neutral tract,

$$cA\,(\theta) = {}^{v}A_{N} + [{}^{c}A - {}^{v}A_{N}]\ W_{C}\,(\theta)$$

where $^{v}A_{N}$ is the neutral tract, ^{c}A the consonant configuration and $W_{C}(\theta)$ another weighting factor. The result of superposition

$$\begin{aligned} cA(t') &= {}^{v}A(t') + {}^{c}A(\theta) - {}^{v}A_{N} \\ &= {}^{v}A(t') + [{}^{c}A - {}^{v}A_{N}]\ W_{C}(\theta). \end{aligned}$$

(Ichikawa and Nakata in a later paper (1968) treat superposition as multiplicative rather than additive.) Nakata and Mitsuoka claim that this rule automatically gives a good approximation of the different shapes of [k] in [ki] and [ko] at the time of maximal constriction: a fact which acoustic systems and earlier articulatory systems handle ad hoc.

The obvious advantage of using shape rules rather than acoustic rules is that the translation from the discrete to the continuous domain becomes more straightforward. The rule for transitions for stops can be stated simply and in the same terms as the rules for glides. But the notion of a target shape is rather unsatisfactory, because only a certain part of the shape is pertinent to any particular phone, and the rest must be arbitrarily specified. In another sense, moreover, a shape model is less interesting than an acoustic system. Ladefoged (1964) has

pointed out that synthesis systems may be classified both as articulatory or acoustic and as parametric or non-parametric. The shape models we have just been discussing are articulatory, but they are not based on any natural parameters comparable to formant frequencies, still less on any set of features. Instead, an arbitrary number of vocal tract cross-sections is used. This is a level of development corresponding to the point in acoustic phonetics when the most significant possible representation of the acoustic spectrum was in terms of a bank of filters. A further limitation of shape systems is that the transitional rules can be little more than arbitrary smoothing rules; it would be very difficult to characterize the changes in shape of the vocal tract differentially segment by segment. In fact, it is a question whether vocal tract shape as such is a significant stage in the speech chain, except in a strictly physical sense. What is needed is a set of parameters for vocal tract shape which would account for the behavior of the tract in the formation of the various sounds and at the same time facilitate a simple statement of rules.

Stevens and House (1955) have suggested a simple three-parameter model for vowel articulation in which the vocal tract is idealized as a tube of varying radius. Two of the parameters are d, the distance of the main constriction from the glottis, and r_0, the radius of the tube at this constriction. The radius r at another point along the tube depends on r_0 and the distance x from the constriction:

$$r - r_0 = .25\,(1.2 - r_0)\,x^2$$

The front portion of the tract (14.5 cm. from the glottis and beyond), however, is characterized by a third parameter $A/2$, the ratio of the area of mouth opening to the length of this position of the tract. This ratio, inversely proportionate to acoustic impedance, varies depending on the protrusion of the lips. These parameters correspond, of course, to the familiar phonetic dimensions of front-back, open-close and rounded-unrounded, and serve to characterize vowels very well. With a static vocal tract analog, Stevens and House were able to use this model to synthesize vowels with the formant frequency ranges observed by Peterson and Barney (1952). These parameters are not, of course, satisfactory for most consonants, if only because the formula for computing r would break down under the circumstances of fricative narrowing and stop closure. Ichikawa et al. (1967) propose another, more general scheme for which the parameters are the maximal constriction point P and the maximum area points V_1 and V_2 of the front and back cavities formed by this constriction. Ichikawa and Nakata (1968) report that they have used this very over-simplified model in a synthesis-by-rule system. It does not seem likely, however, that any parametric description of vocal tract shape will prove satisfactory unless it directly reflects the behavior of the various articulators in some detail. As Ladefoged (1964:208) has observed, 'describing articulations in terms of the highest point of the tongue or the point of maximum constriction of the vocal tract is rather like describing different ways of walking in terms of movements of the big toe or ankle'. But this is as much as to say that it is neces-

sary to go back to the next earlier stage of the speech chain, the stage of articulatory gesture.

4.3 *Articulator Systems*

A number of investigators are attempting to write rules for synthesis in terms of the movements of the individual articulators. The basic approach is to assume a model for the motion of each articulator which is convenient for the statement of the rules. From the states of the articulator models the vocal tract shape, and in turn the acoustic signal, can be determined for a given excitation.

Coker (1967, pers. comm.) uses a modified version of a model suggested by Coker and Fujimura (1966). Two parameters for the lips, one for the velum and four for the tongue determine the shape of the oral vocal tract. The lip parameters indicate the degree of protrusion and of closure; the parameter for the velum indicates its relative elevation; two of the tongue parameters indicate the degree of apical closure and front-back position for the tongue tip; and the other two, the position of the central mass of the tongue in the midsagittal plane. For each phone, target values for these parameters are stored. The stored values are divided into 'important' and 'unimportant'; thus, degree of rounding is unimportant for most sounds but important for [i] and [w] at one extreme and [y] at the other. Interpolation from target to target is accomplished by a 'low pass filter' rule, which produces a certain amount of coarticulation and vowel reduction. The different parameters move at different speeds — for example, the apical parameter is quite fast and the protrusion parameter quite slow. The degree of coarticulation is greater for slowly-moving parameters than for fast ones. Parameter speed is increased in transitions from unimportant to important values and reduced in transitions from important to unimportant values, thus increasing coarticulation for those parameters which specially characterize a particular phone. Parameter timing can also be modified depending on context; this feature of the system is used to provide anticipatory rounding. Target values for each phone are changed simultaneously, except that in a consonant cluster, parameters for different articulators overlap. For each momentary set of articulatory parameter values, the corresponding vocal tract shape is determined, and from the shape, the formant frequencies, which are used to control a resonance synthesizer.

Haggard (Werner and Haggard 1969) has developed a similar model with 11 parameters. Like Coker, he has parameters for lip protrusion and lip closure, for elevation of the velum, and for tongue-tip position and closure. Position, degree of closure, and length of closure are parameters for the body of the tongue, and degree of closure for jaw and glottis. From a momentary description in terms of articulatory parameters, 'construction' (i.e. shape) parameters are derived which describe the vocal tract as a sequence of a few tubes of varying length and cross-sectional area. A nomogram of the sort given by Fant (1960:65) is used to cal-

culate formant frequency values for control of a resonance synthesizer. Target values of articulatory parameters stored for each phone are distinguished as 'marked' or 'unmarked', depending on whether they are characteristic of the articulation of the phone; the distinction is much the same as Coker's important v. unimportant. A further distinction is made between position and closure parameters. The transition of a parameter from the midpoint of one phone to the midpoint of the next is made up of linear segments, and varies depending on the type of each phone (vowel, consonant, or pause), the marking of the two target values, the characteristic rate of each parameter, and the parameter type (position or closure). The rather complex transition rules insure that marked target values for consonant closure parameters will be attained and held, and that progress towards other marked target values will occur over a longer time and at a more rapid rate than toward unmarked values. Thus coarticulation and centralization are provided for. In general a phone can influence only the adjacent phones, but nasalization is provided for by allowing the velar closure parameter to influence several preceding phones.

Henke's (1967) model attempts to handle the same coarticulatory phenomena as Haggard's and Coker's while avoiding a commitment to a parametrization in favor of a naturalistic representation of articulation. Each articulator is represented by a family of 'fleshpoints' on the midsagittal plane. During the motion of an articulator, each point moves along a vector determined by a target location and a target articulator shape. During the early part of its motion, a point first accelerates as the inertia of the articulator is overcome, then attains an appropriate steady velocity, and finally slows as it approaches the target point. The motion of the articulators is determined by a set of attributes stored for each phone. A configurative attribute corresponds to a target location and shape; a strength attribute, to the force which moves an articulator. At any moment, motion may be controlled by attributes associated with one or several successive phones. However, different attributes referring to the same articulatory region cannot both apply at once. A change of attribute will occur at a time dependent on the attributes of the current phone and of the following phone, and upon the progress of articulatory movements determined by other attributes. For example, when articulation of a stop consonant begins, the relevant stop attributes, specifying the shape and location of the articulator and the force of the closure, assume control of the articulator. When closure is attained, the attributes of a following vowel, except those which conflict with the stop attributes, are applied, and attributes of earlier phones are dropped. After the stop is released, all the attributes of the following vowel apply for enough time to allow the articulators to approach the vowel target.

Systems such as those of Coker, Haggard and Henke are more theoretically adequate than shape systems; we are clearly closer to the level of phonetic features. The translation from discrete to continuous domains is more natural because a target is defined for each articulator. We might compare the kind of description

given by these systems to that of an idealized X-ray movie of the vocal tract, from which not only dynamic changes in shape, but also the contribution of each of the individual articulators to the changes in shape are apparent.

4.4 *Neuromotor Command Synthesis*

But the description is still deficient in some respects. The parameters which describe articulatory motion may seem the obvious ones, and may be empirically successful, but they have no necessary theoretical basis. The various articulators, of course, are not free to move at random but only to and from a limited number of targets. This limitation on the number of targets accounts for the limited number of values which can be assumed even by features associated with such a complex articulator as the tongue. If we could go a stage further back in the speech chain, and synthesize speech at the level of the neuromotor commands which control the muscles of the articulators, we might be able to account for these significant limitations. Fortunately, electromyographic techniques can help us here (e.g. Harris et al. 1965; Fromkin 1966). From measurements of the voltages picked up during speech by electrodes placed in the vocal tract, it is possible to make some plausible inferences about muscle activity — and hence about the corresponding neuromotor commands — in the production of the sounds of speech.

The synthetic counterpart of electromyographic analysis would describe speech in terms of a series of commands to the muscles of the vocal tract. An approach to this kind of synthesis has been made by Hiki, who has developed a description of jaw and lip movement using muscle parameters (Hiki and Harshman 1969). The forward part of the vocal tract is treated as an acoustic tube of varying length, height and width. The value for each dimension depends on the positive or negative force exerted by lip and jaw muscles, and each of these muscles may affect other dimensions as well. Muscles of the lips which affect the same dimensions in the same way are grouped together, and the same is the case for muscles of the jaw. The force exerted by such a group of muscles (actually the effect of several neuromotor commands) is a parameter of the system. Four lip and two jaw parameters are used. The forces acting separately on lip and jaw are combined to produce a description of shape, and with this partial model, labial sounds can be synthesized with a vocal tract analog. More recently Hiki has extended his investigations to the tongue (Hiki 1970).

Clearly, synthesis by rule must move in the direction suggested by Hiki's work. Only with models of this sort, making use of the earliest observable stage of the speech chain, will it be possible to gain insight into the nature of individual gestures and their relative timing. It is significant, however, that myographic synthesis, as represented by Hiki's scheme, seems to lead to an increase rather than a decrease in the number of parameters, as compared with articulatory models, even though

several muscles exerting parallel forces are grouped under one parameter. Though the neuromotor commands for e.g. lip closure are similar for the different manner classes of labial sounds (Harris et al. 1965), the relationship between this gesture and the neuromotor commands which produce it is not a simple one. This suggests that the connection between the phonetic feature corresponding to lip closure and the neuromotor commands may not be simple, either; perhaps the realization of some value of a phonetic feature as a unitary psychological gesture may actually involve a complex neuromotor program. This view is reinforced by the recent finding of MacNeilage and DeClerk (1969) that coarticulation appears even in electromyographic data.

4.5 *Synthesis of Excitational and Prosodic Features*

Our discussion so far has been concerned with the synthesis of segmental phones and with supraglottal articulation and its acoustic consequences. A synthesis by rule scheme also has to take into account excitational, prosodic and demarcative features, the associated glottal and subglottal events, and the acoustic correlates of these events.

Both resonance and vocal tract analog synthesizers provide periodic and noisy excitation sources, periodic excitation being used for vowels, sonorants and voiced stops; noisy excitation for [h], aspiration, and frication. In resonance synthesizers, separate circuits (either fixed filters or variable frequency resonators) are ordinarily provided for shaping high-frequency frication; in vocal tract analog synthesizers, noise is inserted at various segments in the tract, depending on the place of articulation of the fricative. With such facilities the different kinds of excitation are readily simulated; the only problem is to write rules for the changes from one excitation source to another. This aspect of synthesis by rule has not been taken very seriously; usually the duration of the excitation appropriate for a phone is identical with the nominal duration of the phone itself. In the case of voiceless stops, however, this approach requires including part of the transition to the following vowel in the stop, as was done by Holmes et al. (1964). Another solution is to specify, as a characteristic of the voiceless consonant, the appropriate amount of devoicing of the following phone, as we have done (Mattingly 1968a). What is really required, however, is a rule specifying voice-onset time negatively or positively relative to the instant of release, as the work of Lisker and Abramson (1967) suggests. For medial and final voiced consonant and consonant clusters, increased duration of the preceding vowel is well known to be an important cue (Kenyon 1950:63; Denes 1955) and some systems have taken account of it, e.g. Mattingly (1968a), Rabiner (1969).

Rather more attention has been given to prosodic and demarcative features such as stress, accent, intonation, juncture and pause, which interact with inherent prop-

erties of a phone to determine duration, fundamental frequency and intensity.

In our own prosodic control scheme for British English (Mattingly 1966), two degrees of stress and three common intonation contours (fall, fall-rise and rise) can be marked in the input. The F_0 rules specify a falling contour during the 'head' of the breath group; the slope varies with the quality of the syllable nucleus. Voiceless consonants cause a 'pitch skip'; stressed syllables, a smooth rise in F_0. The required terminal intonation contour is imposed on the 'tail' — the last stressed syllable and any following syllables. Each possible syllable nucleus has an inherent duration which is increased multiplicatively by stress. In prepausal syllables, the duration of all phones is increased and amplitude is gradually diminished. More recently (Mattingly 1968a), we have used similar rules for synthesis of General American, and in addition, provided for the durational effects of juncture.

Rabiner (1969) follows the model proposed by Lieberman (1967) in which the overall fundamental contour is determined by subglottal air pressure, except for the so-called 'marked' breath group where laryngeal tensing produces a terminal rising contour. Four degrees of stress can be indicated; the higher the stress, the greater the increase in F_0 on the stressed syllable. Duration increases additively with the openness and tenseness of the vowel and the degree of stress, as well as being affected by the following consonant.

Hiki and Oizumi (1967) have developed prosodic rules similar to those of Rabiner and Mattingly. The F_0 rules deal with pitch accent, emphasis, terminal contours, the overall contour and individual differences; the duration rules take into account the inherent duration of phones, pause length and accent; and interestingly, the effect of changes of tempo on these features.

Umeda et al. (1968) determine prosodic patterns for English directly from ordinary printed text and use them to control the vocal-tract synthesis by rule scheme of Matsui (1968). Syntactic rules of a primitive kind are used to divide an utterance into blocks corresponding to breath groups. An overall F_0 contour is imposed on each breath group. Word stress (along with the phonetic transcription for a word) is determined by table lookup, and the fundamental frequency and duration are increased accordingly. Intonation contours and pause duration are derived from the punctuation, if any, following each block.

Vanderslice (1968) has also considered prosodic feature from the standpoint of the problems involved in the conversion of orthographic text to sound, correctly distinguishing some features not included in earlier systems and proposing rules for their synthesis. Two degrees of pitch prominence are used instead of one: 'accent' and 'emphasis', the latter for contrastive and emphatic stress. The features 'cadence' and 'endglide' replace the traditional fall, fall-rise and rise, cadence being equivalent to a fall, end-glide to a rise, and cadence followed by end-glide to fall-rise. 'Pause' is a separate feature. To account for the raising of F_0 in quoted material and its lowering in parenthetical material, the features 'upshift' and 'downshift' respectively are used. An additional group of 'indexical' features is proposed

for stylistic variation, e.g. 'dip', for the downward pitch prominence noted by Bolinger (1958).

While only a few terminal intonation contours carry grammatical information and are required for synthesis of ordinary discourse, many more occur in colloquial speech. Iles (1967), following a scheme of tones (i.e. terminal contours) proposed by Halliday (1963), has attempted to synthesize some of these tones, imposing the contours on segmental synthetic speech generated by PAT in the manner of Holmes et al. (1964).

The models of the behavior of F_0 discussed so far assume a basic contour for the whole breath group, on which stress and terminal intonation contours are imposed — an approach consistent both with the work of British students of intonation from Armstrong and Ward (1931) to O'Connor and Arnold (1961), and with the 'archetypal' model of intonation proposed by Lieberman (1967). Another possible approach is to characterize a breath group as a series of pitch-levels, as in Pike's (1945) well-known scheme. Shoup (pers. comm.) has synthesized sentences in which the F_0 contours corresponding to such descriptions are realized, taking into account the stress, the vowel quality, and the excitation of the preceding consonant and the frequency at the beginning of the syllable.

Synthesis by rule of prosodic features has come to receive serious attention only quite recently, by comparison with synthesis of segmental features. We have only just begun to understand what is easy, what is difficult; what is relevant and what is irrelevant. Of the three major correlates of the prosodic features, intensity has proved the least sensitive and the least important. F_0 has attracted the most interest: considerable success has been attained in producing convincing stress and terminal intonation contours by rule, and the articulatory mechanism has been simulated. Duration, however, remains a serious problem. No one has yet produced even an empirically successful set of duration rules, and it is far from clear what theoretically adequate rules would be like. Presumably there are some durational effects which are really automatic consequences of the articulation: formant transition durations surely fall into this category. A second group of effects are truly temporal, but subject to phonological rule: vowel length, for example. Finally, there are effects which are, to some extent, under the conscious control of the speaker: speaking rate, for instance. All these different effects are superimposed in actual speech: sorting them out is a major task for synthesis by rule.

In the prosodic schemes we have just been describing, even those which model supraglottal shape or articulatory movement, the prosodic features are still being simulated purely acoustically. No attempt is made to model explicitly the articulatory mechanisms which are responsible for the variation in the acoustic correlates. Unfortunately, the prosodic articulatory mechanisms are much less well understood than those which underlie segmental features, which explains in part the lack of unanimity concerning the appropriate treatment of prosodic features at the phonological level.

Flanagan, however, has made impressive progress with his computer simulations of vocal tract excitation (Flanagan and Landgraf 1968; Flanagan and Cherry 1969). Voicing is represented by the output of a system consisting of two masses, corresponding to the vocal cords, oscillating so as to vary the cross-sectional area of the passage between them, corresponding to the glottis. At one end of the passage is a source of air varying in pressure, representing the lungs; at the other end is a vocal tract analog. In response to the subglottal air pressure, the displacement of each mass increases, as does the air flow through the glottis. But this increase in air flow results in an increase in negative Bernouilli pressure between the two masses, reducing the displacement, so that oscillation occurs. The frequency of the oscillation varies with the subglottal pressure, with the size of the two masses and their stiffness (vocal-cord tension), and with the acoustic impedance of the vocal tract analog, which depends on the phone being synthesized. Thus the model allows simulation of the separate roles of lung pressure and cord tension in determining F_0 and takes account of interaction with the supraglottal tract. The same model serves to simulate frication, which occurs at a constriction in the supraglottal tract when the constriction is sufficiently narrow and the pressure behind sufficiently great. When both glottal and fricative excitation are present, as in a voiced fricative, the pattern of pitch-synchronous bursts in the noise is simulated by the model.

4.6 *Synthesis Using Phonological Rules*

As we have just seen, a substantial amount of research effort in speech synthesis by rule has been concerned with what we have called phonetic capacity. Other components — phonetic skill, phonological competence and capacity — have received relatively little attention. There are various reasons for this. The quality of speech synthesized by rule has only quite recently been good enough to serve as a vehicle for research in these other components; moreover, many of those engaged in synthesis by rule have been content to operate with a fairly rough and ready view of phonology, because they are more interested in the physical aspects of speech, either acoustic or articulatory, than with phonetic capacity as such, or its relationship to other components. A few years ago this might have mattered much less; but recent impressive developments in generative phonology make it important that synthesis by rule display greater sophistication in this area if linguists are to take it seriously.

A few scattered efforts have been made. In our own work (Mattingly 1968a), we have drawn a distinction between the synthesis by rule program with the associated hardware, representing the universal aspects of speech (phonological and phonetic capacity) and the rules of a particular language or dialect (phonological competence) which were an input to the program. In practice this distinction is not made consistently: certain matters are handled in the rules which more properly belong in the program, and conversely.

The system also provides a kind of primitive phonological framework, in that it allows the statement of ordered, contrast-dependent allophone rules which modify the stored data for synthesis of a phone. The description of the contexts in which a rule can be altered are built up from a limited set of binary contextual features, e.g. 'prevocalic', 'post-vocalic', 'stressed': these contextual features are part of the program. Thus it is claimed that the nature of phonological capacity is such that only a small fraction of the conceivable contexts in fact occur in the phonological rules of natural languages — a claim with which Chomsky and Halle (1968 : 400–1) appear to be sympathetic. The program has been used to synthesize both the General American and the Southern British dialects of English (Haggard and Mattingly 1968).

In this system the prosodic rules are phonetic: phonologically predictable stress and intonation effects must be marked in the input. Vanderslice (1968), however, has proposed a set of rules for predicting the occurrence, in English, of the prosodic features for which his definitions have been given above, in particular accent. His strategy is to assign provisional accents to all lexically-stressed syllables and then to delete certain of these accents. For example his 'rhythm rule' deletes the middle one of three consecutive accentable syllables in the same sense group. If a word such as 'unknown' is assumed to have two accentable syllables in its lexical form, this rule accounts nicely for the shifting stress in such words. Other rules proposed by Vanderslice rely on syntactic or semantic conditions: these conditions will somehow have to be marked at the input to the phonology.

Finally, mention should be made of Allen's (1968) programming of the Chomsky and Halle (1968) rules for the assignment of accent. At this writing, so far as we know, no one has thus far attempted a program for synthesis of English based on the Chomsky-Halle rules for segmental phonology, or even a computer simulation with phonetic-feature matrices as output. Fromkin and Rice (1970), however, have developed a program for which the input format follows closely that of the Chomsky-Halle phonological conventions, and permits the testing of a set of phonological rules.

5. SUMMARY AND CONCLUSIONS

We must now try to sum up the current state of synthesis by rule and to indicate the directions the work may be expected to take in the future. As we have seen, it is possible to synthesize speech by rule which is not only intelligible but also reasonably acceptable to a native speaker. Moreover, the trend of research in the past few years has been toward the development of systems of increasing phonetic sophistication with correspondingly greater theoretical interest. In the synthesis of segmental sounds, the emphasis has shifted from acoustic models to vocal-tract shape models, and from shape models to articulator models; a similar trend is

evident in prosodic synthesis. Though much of this work is motivated, in the first instance, by an interest in the physical aspects of speech production, it is clearly also leading toward an increased understanding of phonetic capacity.

In the future, it is desirable, first of all, that synthesis by rule become a tool for the study of phonological capacity and competence. In practice, this would mean the development, according to the principles of modern generative grammar, of phonological descriptions of languages, the outputs of which would be converted to speech by a synthesis system. Such enterprises would be valuable for two reasons: first, they would tend to correct the present rather off-hand conceptions of phonology entertained by many of those who are now doing synthesis by rule. Second, they would, on the other hand, compel the phonologist to relate his descriptive rules to some clearly-defined concept of phonetic capacity, and permit him to test utterances produced by his phonological rules against the intuitions of the native speaker. When the generative grammarian is doing syntax, it is quite natural for him to offer examples in a form such that any native speaker can determine their grammaticality; the grammarian should be able to operate on the same basis when he is doing phonology, and he can, if he uses synthesis by rule. It would also seem desirable to increase considerably the number of dialects and languages for which rules for synthesis have been written. Most of the work thus far has been done in English and Japanese; many other languages should be synthesized as part of a general effort to explore the different versions of phonological competence.

The use of synthesis by rule to study phonological competence and capacity will, of course, compel attention to the central problem of phonetic capacity, that of enumerating and defining the universal set of phonetic features, and in the process giving increased psychological meaning to the notion 'feature'. This means, in practice, the development of systems in which phonetic feature matrices are the input to the part of the program which simulates phonetic capacity. Though various feature-like entities have played a part in several of the systems we have discussed, none of these systems really represents a consistent attempt to synthesize speech using what the phonologist would regard as phonetic features. Many problems must still be worked out. For example, a feature involving a particular articulator can be equated with a gesture of the articulator toward a particular target, but the synthesis by rule system must somehow define just what it is that accounts for the psychological unity of this gesture, regardless of the original position of the articulator. For manner features, the problem is still more acute: the system must explain how features such as 'stop-continuant' can be given a plausible unitary definition in terms of phonetic capacity, even though physically quite different articulations may be used for the production of the various stops and continuants. If the feature is defined in part by feedback of some kind, this must be part of the synthesis system.

Nor is the matter of the translation from the discrete to the continuous as yet

handled really adequately by the systems we have discussed. Having recognized that the targets for each articulator must be separately described, we must now try to account in some principled way for the coordination of the movements of the various articulators towards their targets. Present systems, in which each phone is dealt with in turn, and is affected by the preceding and following phone but no others (with the exception of arrangements in some systems to nasalize several preceding phones), are too restrictive to account for the fact that coarticulation may extend over several phones (Kozhevnikov and Chistovich 1965). The assumption of these systems is that the changing of targets for the various articulators is synchronized phone by phone — an assumption which works empirically after a fashion, but masks the real problem of how and to what extent the movements of articulators *are* synchronized, and how much account phonetic capacity must take of the synchronization process. It has frequently been suggested that the syllable, which certainly seems to have psychological reality — and therefore some role in phonetic capacity — is the unit of coarticulation. Clearly there is a need of a synthesis by rule system which explores this possibility.

Another area which needs a great deal of further attention is the nature of the demarcation between phonetic capacity and phonological competence. We want to reflect this separation as clearly as possible in a synthesis by rule system; unfortunately, it is not always possible to distinguish in particular cases between 'intrinsic' allophones (belonging to phonetic capacity) and 'extrinsic' allophones (belonging to phonological competence) (Wang and Fillmore 1961). Moreover, Tatham (1969b) has argued that since such an 'intrinsic' difference as front v. back [k] can be distinctive in some languages, we have to provide for countermanding, in special cases, of a normal rule of phonetic capacity by phonological competence. This is actually, as Tatham points out, a problem relating to 'markedness', an issue involving the relationship between the two components which has concerned phonologists from Troubetskoy (1939:79) and the Prague School to Chomsky and Halle (1946: 402ff.).

Finally, we can also look forward to increasing our understanding of the elusive matter of phonetic skill through synthesis by rule. The systems we have discussed all assume an ideal or at least typical speaker with a consistent style; questions of phonetic skill are avoided. But given some reasonably satisfactory representation of other components, we can begin to derive auxiliary sets of rules representing phonetic skill, and consisting of a series of modifications to the parts of the system representing phonological competence and phonetic capacity. Suppose, for instance, that we wish to investigate the productive and perceptual factors of speaker variation. These are matters in part of the physical characteristics of the speaker (and so will involve adjustments of the synthesizer itself) but also of phonetic skill. At present the preferred methods of study are subjective ratings of speakers and examination of spectrograms. But it should be possible, using synthesis by rule, to try to mimic speakers and to study listeners' perceptions of such mimicry under

quite specific assumptions about the speaker's and the listeners' phonetic and phonological capacity and the rules of their language.

Questions such as these make it apparent that synthesis by rule forces attention to precisely those phonetic problems which are fundamental to phonology. We hope that some phonologists will be sufficiently intrigued to join in the search for the answers.

REFERENCES

ABRAMSON, ARTHUR S., and LEIGH LISKER. 1970. Laryngeal behavior, the speech signal and phonological simplicity. PICL 10/4.123–29.

ALLEN, JONATHAN. 1968. A study of the specification of prosodic features of speech from a grammatical analysis of printed text. Unpublished Ph.D. Dissertation, Massachusetts Institute of Technology.

ANTHONY, JAMES. 1964. True model of the vocal tract. JAcS 36.1037. (A)

ANTHONY, JAMES, and WALTER LAWRENCE. 1962. A resonance analogue speech synthesizer. Proceedings of the Fourth International Congress on Acoustics, ed. by A. Kjerbya Nielsen, Paper G43. Copenhagen, Organization Committee, Fourth ICA.

ARMSTRONG, LILIAS C., and IDA C. WARD. 1931. A handbook of English intonation. 2nd. ed. Cambridge, Heffer.

ATKINSON, R. C., and H. A. WILSON. 1968. Computer-assisted instruction. Science 162.73–7.

BAXTER, BRENT, and WILLIAM J. STRONG. 1969. WINDBAG — a vocal-tract analog speech synthesizer. JAcS 45.309. (A)

BOLINGER, DWIGHT. 1958. A theory of pitch accent in English. Word 14.109–49.

BORST, JOHN M. 1956. Use of spectrograms for speech analysis and synthesis. JAudEngSoc 4.13–23.

CHOMSKY, NOAM. 1965. Aspects of the theory of syntax. Cambridge, Mass. MIT Press.

———. 1968. Language and mind. New York, Harcourt.

CHOMSKY, NOAM, and MORRIS HALLE. 1968. The sound pattern of English. New York, Harper.

COKER, CECIL H. 1965. Real-time formant vocoder, using a filter bank, a general-purpose digital computer, and an analog synthesizer. JAcS 38.940. (A)

———. 1967. Synthesis by rule from articulatory parameters. Conference preprints, 1967 Conference on Speech Communication and Processing, pp. 53–3. Bedford, Mass., Air Force Cambridge Research Laboratories.

COKER, CECIL H., and PETER CUMMISKEY. 1965. On-line computer control of a formant synthesizer. JAcS 38.940. (A)

COKER, CECIL H., and OSAMU FUJIMURA. 1966. Model for specification of vocal-tract area function. JAcS 40.1271. (A)

COOPER, FRANKLIN S. 1950. Spectrum analysis. JAcS 22.761–2.

——. 1962. Speech synthesizers. PICPS 4.3–13.

COOPER, FRANKLIN S., JANE H. GAITENBY, IGNATIUS G. MATTINGLY and NORIKO UMEDA. 1969. Reading aids for the blind: A special case of machine-to-man communication. IEEE Trans. Audio 17.266–70.

DENES, PETER B. 1955. Effect of duration on the perception of voicing. JAcS 27.761–64.

——. 1965. 'One-line' computing in speech research. JAcS 38.934. (A)

——. 1970. Some experiments with computer synthesized speech. Behav. Res. Meth. & Instru. 2.1–5.

DENNIS, JACK B. 1963. Computer control of an analog vocal tract. JAcS 35.1115. (A)

DENNIS, JACK B., EDWARD C. WHITMAN, and RAYMOND S. TOMLINSON. 1964. On the construction of a dynamic vocal-tract model. JAcS 36.1038.

DIXON, N. REX, and H. DAVID MAXEY. 1968. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. IEEE Trans. Audio 16.40–50.

——. 1970. Functional characteristics of an on-line, computer-controlled speech synthesizer. JAcS 47.93. (A)

DUDLEY, HOMER. 1939. The Vocoder. Bell Labs. Rec. 18.122–6.

DUDLEY, HOMER, R. R. RIESZ, and S. S. A. WATKINS. 1939. A synthetic speaker. J. Franklin Inst. 227.739–64. Philadelphia.

DUDLEY, HOMER, and THOMAS H. TARNOCZY. 1950. The speaking machine of Wolfgang von Kempelen. JAcS 22.151–66.

DUNN, HUGH K. 1950. Calculation of vowel resonances, and an electrical vocal tract. JAcS 22.740–53.

EIMAS, PETER, EINAR R. SIQUELAND, PETER JUSCZYK, and JAMES VIGORITO. 1971. Speech perception in infants. Science 171.303–06.

ESTES, S. E., H. R. KERBY, H. DAVID MAXEY, and ROBERT M. WALKER. 1964. Speech synthesis from stored data. IBM J. 8.2–12.

FANT, C. GUNNAR M. 1956. On the predictability of formant levels and spectrum envelopes from formant frequencies. For Roman Jakobson, ed. by Morris Halle et al., pp. 109–20. The Hague, Mouton.

——. 1958. Modern instruments and methods for acoustic studies of speech. PICL 8.282–358.

——. 1960. Acoustic theory of speech production. The Hague, Mouton.

FANT, C. GUNNAR M., JANOS MARTONY, ULF RENGMAN, and ARNE RISBERG. 1963. OVE II synthesis strategy. Proceedings of the Speech Communications Seminar, paper F5. Stockholm, Speech Transmission Laboratory, Royal Institute of Technology.

FLANAGAN, JAMES L. 1957. Note on the design of 'terminal' analog speech synthesizers. JAcS 25.306–10.

——. 1965. Speech analysis, synthesis and perception. Berlin, Springer.

FLANAGAN, JAMES L., and L. CHERRY. 1969. Excitation of vocal-tract synthesizers. JAcS 45.764–9.

FLANAGAN, JAMES L., CECIL H. COKER, and CAROL M. BIRD. 1962. Computer simulation of a formant vocoder synthesizer. JAcS 34.2003. (A)

FLANAGAN, JAMES L., and LORINDA L. LANDGRAF. 1968. Self-oscillating source for vocal-tract synthesizers. IEEE Trans. Audio 16.57–64.

FOURCIN, ADRIAN. 1960. Potential dividing function generator for the control of speech synthesis. JAcS 32.1501. (A)

FROMKIN, VICTORIA A. 1966. Neuromuscular specification of linguistic units. L&S 9.170–99.

FROMKIN, VICTORIA A., and D. LLOYD RICE. 1970. An interactive phonological rule testing system. WPP 14.8.

GARIEL, [?]. 1879. Machine parlante de M. Faber. J. Physique Théorique et Appliquée 8.274–5.

GLACE, DONALD A. 1968. Parallel resonance synthesizer for speech research. JAcS 44.391. (A)

HAGGARD, MARK P., and IGNATIUS G. MATTINGLY. 1968. A simple program for synthesizing British English. IEEE Trans. 16.95–9.

HALLIDAY, MICHAEL A. K. 1963. The tones of English. ArchL 15.1–28.

HARRIS, CYRIL M. 1953. A study of the building blocks of speech. JAcS 25.962–9.

HARRIS, KATHERINE S., GLORIA F. LYSAUGHT, and MALCOLM M. SCHVEY. 1965. Some aspects of the production of oral and nasal labial stops. L&S 8.135–47.

HEFFNER, ROE-MERRILL. 1950. General phonetics. Madison, Univ. of Wisconsin Press.

HENKE, WILLIAM. 1967. Preliminaries to speech synthesis based upon an articulatory model. Conference preprints, 1967 Conference on Speech Communication and Processing, pp. 170–7. Bedford, Mass., Air Force Cambridge Research Laboratories.

HIKI, SHIZUO. 1970. Control rule of the tongue movement for dynamic analog speech synthesis. JAcS 47.85. (A)

HIKI, SHIZUO, and RICHARD HARSHMAN. 1969. Speech synthesis by rules with physiological parameters. JAcS 46.111. (A)

HIKI, SHIZUO, and JURO OIZUMI. 1967. Controlling rules of prosodic features for continuous speech synthesis. Conference preprints, 1967 Conference on Speech Communication and Processing, pp. 23–6. Bedford, Mass., Air Force Cambridge Research Laboratories.

HIKI, SHIZUO, RICK RATCLIFFE, STAN HUBLER, and PETE METEVELIS. 1968. Notes on LASS circuitry. WPP 10.12–41.

HOLMES, JOHN N. 1961. Research on speech synthesis carried out during a visit to the Royal Institute of Technology, Stockholm, from November, 1960 to March, 1961. Report JU 11–4. Eastcote (England), Joint Speech Research Unit.

HOLMES, JOHN N., IGNATIUS G. MATTINGLY, and JOHN N. SHEARME. 1964. Speech synthesis by rule. L&S 7.127–43.

HONDA, TAKASHI, SEIICHI INOUE, and YASUO OGAWA. 1968. A hybrid control system of a human vocal tract simulator. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, pp. 175–8. Tokyo, International Council of Scientific Unions.

ICHIKAWA, AKIRA, YASUAKI NAKANO, and KAZUO NAKATA. 1967. Control rule of vocal-tract configuration. JAcS 42.1163. (A)

ICHIKAWA, AKIRA, and KAZUO NAKATA. 1968. Speech synthesis by rule. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, pp. 171–4. Tokyo, International Council of Scientific Unions.

ILES, LAURENCE A. 1967. M. A. K. Halliday's "Tones of English" in synthetic speech. Work in Progress 1.24–6. Edinburgh, Department of Phonetics, Edinburgh University.

——. 1969. Speech synthesis by rule. Work in Progress 3.23–5. Edinburgh, Department of Phonetics and Linguistics, Edinburgh University.

INGEMANN, FRANCES. 1957. Speech synthesis by rule. JAcS 29.1255. (A)

——. 1960. Eight-parameter speech synthesis. JAcS 32.1501. (A)

KACPROWSKI, JANUSZ, and WLADYSLAW MIKIEL. 1968. Recent experiments in parametric synthesis of Polish speech sounds. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, pp. 191–4. Tokyo, International Council of Scientific Unions.

KATO, YASUO, KAZUO OCHIAI, and SHINTARO AZAMI. 1968. Speech synthesis by rule supplementarily using natural speech segments. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, pp. 199–202. Tokyo, International Council of Scientific Unions.

KELLY, JOHN L., and LOUIS J. GERSTMAN. 1961. An artificial talker driven from a phonetic input. JAcS 33.835. (A)

KELLY, JOHN L., and CAROL LOCHBAUM. 1962. Speech synthesis. Proceedings of the Speech Communications Seminar, paper F7. Stockholm, Speech Transmission Laboratory, Royal Institute of Technology.

KEMPELEN, WOLFGANG R. VON. 1791. Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine. Wien, J. B. Degen.

KENYON, JOHN S. 1950. American pronunciation. 10th ed. Ann Arbor, Mich., Wahr.

KIM, CHIN-WOO. 1966. The linguistic specification of speech. WPP 5.

KOENIG, W. H., H. K. DUNN, and L. Y. LACEY. 1946. The sound spectrograph. JAcS 18.19–49.

KOZHEVNIKOV, V. A., and LUDMILA A. CHISTOVICH. 1965. Rech' artikuliatsia i vospriiatie. Translated (1966) as Speech: Articulation and perception. Washington, D.C., Joint Publications Research Service.

LADEFOGED, PETER. 1964. Some possibilities in speech synthesis. L&S 7.205–14.

——. 1967. Linguistic phonetics. WPP 6.

LAWRENCE, WALTER. 1953. The synthesis of speech from signals which have a low information rate. Communication theory, ed. by Willis Jackson, pp. 460–71. London, Butterworth.

LIBERMAN, ALVIN M., FRANKLIN S. COOPER, DONALD P. SHANKWELIER, and MICHAEL STUDDERT-KENNEDY. 1967. Perception of the speech code. PsychRev 74.431–61.

LIBERMAN, ALVIN M., KATHERINE S. HARRIS, HAROLD S. HOFFMAN, and BELVER C. GRIFFITH. 1957. The discrimination of speech sounds within and across phoneme boundaries. JExPsych 53.358–68.

LIBERMAN, ALVIN M., KATHERINE S. HARRIS, JO ANN KINNEY, and HARLAN LANE. 1961. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. JExPsych 61.379.88.

LIBERMAN, ALVIN M., FRANCES INGEMANN, LEIGH LISKER, PIERRE C. DELATTRE, and FRANKLIN S. COOPER. 1959. Minimal rules for synthesizing speech. JAcS 31.1490–9.

LIEBERMAN, PHILIP. 1967. Intonation, perception and language. Cambridge, Mass., MIT Press.

LILJENCRANTS, JOHAN C. W. A. 1968. The OVE III speech synthesizer. IEEE Trans. Audio 16.137–40.

LINDBLOM, BJÖRN. 1963. Spectrographic study of vowel reduction. JAcS 35. 1773–81.

LISKER, LEIGH, and ARTHUR S. ABRAMSON. 1967. Some effects of context on voice onset time in English stops. L&S 10.1–28.

LISKER, LEIGH, FRANKLIN S. COOPER, and ALVIN M. LIBERMAN. 1962. The uses of experiment in language description. Word 18.82–106.

MACNEILAGE, PETER F., and JOSEPH L. DECLERK. 1969. On the motor control of coarticulation of CVC monosyllables. JAcS 45.1217–33.

MATSUI, EIICHI. 1968. Computer-simulated vocal organs. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, pp. 151–4. Tokyo, International Council of Scientific Unions.

MATTINGLY, IGNATIUS G. 1966. Synthesis by rule of prosodic features. L&S 9.1–13.

——. 1968a. Synthesis by rule of General American English. Supplement to Status Report on Speech Research. New York, Haskins Laboratories.

——. 1968b. Experimental methods for speech synthesis by rule. IEEE Trans. Audio 16.198–202.

——. 1971. Synthesis by rule as a tool for phonological research. L&S 14.47–56.

MATTINGLY, IGNATIUS G., and ALVIN M. LIBERMAN. 1969. The speech code and the physiology of language. Information processing and the nervous system, ed. by K. N. Leibovic, pp. 97–117. Berlin, Springer.

MERMELSTEIN, PAUL. In press. Computer simulation of articulatory activity in

speech production. Proceedings of the International Joint Conference on Artificial Intelligence, Washington, D.C., 1969, ed. by D. E. Walker and L. M. Norton. New York, Gordon & Breach.

MOFFIT, ALAN R. 1969. Speech perception by 20–24 week old infants. Paper presented to the Society for Research in Child Development, Santa Monica, Calif., March, 1969.

MUNSON, W. A., and H. C. MONTGOMERY. 1950. A speech analyzer and synthesizer. JAcS 22.678. (A)

NAKATA, KAZUO, and T. MITSUOKA. 1965. Phonemic transformation and control aspects of synthesis of connected speech. J. Radio Res. Labs. 12.171–86.

O'CONNOR, J. DESMOND, and G. F. ARNOLD. 1961. Intonation of colloquial English. London, Longmans.

ÖHMAN, SVEN E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. JAcS 39.151–68.

——. 1967. Numerical model of coarticulation. JAcS 41.310–20.

OIZUMI, JURO, SHIZUO HIKI, and YOSHINARI KANAMORI. 1967. Continuous speech synthesis from phonemic symbol input. Reports of the Research Institute of Electrical Communication 19.241–5. Sendai, Japan, Tohoku University.

PAGET, RICHARD. 1930. Human speech. London, Routledge and Kegan Paul.

PETERSON, GORDON E., and H. L. BARNEY. 1952. Control methods used in a study of the vowels. JAcS 24.175–84.

PETERSON, GORDON E., WILLIAM S.-Y. WANG, and EVA SIVERTSEN. 1958. Segmentation techniques in speech synthesis. JAcS 30.739–42.

PIKE, KENNETH L. 1945. The intonation of American English. Ann Arbor, Mich., University of Michigan Press.

RABINER, LAWRENCE R. 1967. Speech synthesis by rule: An acoustic domain approach. Bell System Tech. J. 47.17–37.

——. 1968. Digital-formant synthesizer for speech-synthesis studies. JAcS 43.822–8.

——. 1969. A model for synthesizing speech by rule. IEEE Trans. Audio 17.7–13.

RAO, P. V. S., and R. B. THOSAR. 1967. SPEECH: a software tool for speech synthesis experiments. Technical Report 38. Bombay, Tata Institute for Fundamental Research.

ROSEN, GEORGE. 1958. Dynamic analog speech synthesizer. JAcS 30.201–9.

SAITO, SHIZO, and SHIN'ICHIRO HASHIMOTO. 1968. Speech synthesis system based on interphoneme transition unit. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, pp. 195–8. Tokyo, International Council of Scientific Unions.

SCOTT, ROBERT J., DONALD M. GLACE, and IGNATIUS G. MATTINGLY. 1966. A computer-controlled on-line speech synthesizer system. Digest of technical

papers, 1966 IEEE International Communications Conference, pp. 104–5. Philadelphia, IEEE.

SHANKWEILER, DONALD, and MICHAEL STUDDERT-KENNEDY. 1967. Identification of consonants and vowels presented to left and right ears. QJEP 19.59–63.

SHEARME, JOHN N., and JOHN N. HOLMES. 1962. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1–formant 2 plane. PICPS 4.234–40.

STEVENS, KENNETH N., and ARTHUR S. HOUSE. 1955. Development of a quantitative description of vowel articulation. JAcS 27.484–93.

STEVENS, KENNETH N., S. KASOWSKI, and C. GUNNAR M. FANT. 1953. An electrical analog of the vocal tract. JAcS 25.734–42.

STOWE, ARTHUR N., and D. B. HAMPTON. 1961. Speech synthesis with pre-recorded syllables and words. JAcS 33.810–1.

TATHAM, MARCEL A. A. 1969a. Experimental phonetics and phonology. Occasional Papers 5.14–9. Colchester, University of Essex Language Centre.

——. 1969b. On the relationship between experimental phonetics and phonology. Occasional Papers 5.20–6. Colchester, University of Essex Language Centre.

TOMLINSON, RAYMOND S. 1965. SPASS — an improved terminal-analog speech synthesizer. JAcS 38.940. (A)

TROUBETZOY, NICOLAS S. 1939. Principes de phonologie. Tr. 1949 J. Cantineau. Paris, Klincksieck.

UMEDA, NORIKO, EIICHI MATSUI, TORAZO SUZUKI, and HIROSHI OMURA. 1968. Synthesis of fairy tales using an analog vocal tract. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B159–62. Tokyo, International Council of Scientific Unions.

VANDERSLICE, RALPH. 1968. Synthetic elocution. WPP 8.

WANG, WILLIAM S.-Y., and CHARLES J. FILLMORE. 1961. Intrinsic cues and consonant perception. JSHR 4.130–6.

WERNER, EDWENNA, and MARK HAGGARD. 1969. Articulatory synthesis by rule. Speech synthesis and perception, Progress Report 1.1–35. Cambridge, Psychological Laboratory, University of Cambridge.

WHEATSTONE, CHARLES. 1837. Review of On the vowel sounds, by Robert Willis, Le mécanisme de parole, by Wolfgang v. Kempelen (= v. Kempelen, 1791), and Tentamen coronatum de voce, by C. G. Kratzenstein. London & Westminster Rev. 6 and 28.27–41.

YOUNG, ROBERT W. 1948. Review of U.S. Patent 2,432,123, Translation of visual symbols, R. K. Potter, assignor (9 December 1947). JAcS 20.888–9.