Journal of Phonetics (1979) 7, 279–312

Speech perception: a model of acoustic–phonetic analysis and lexical access

Dennis H. Klatt

Room 36-523, *Massachusetts Institute of Technology, Cambridge, MA* 02139, *U.S.A.*

Received 1 *November* 1978

Abstract: Lexical hypothesis formation from acoustic input is an important component of the normal speech perception process. Any model of bottom-up lexical access must address the well-known problems of (1) acoustic-phonetic non-invariance, (2) phonetic segmentation, (3) time normalization, (4) talker normalization, (5) specification of lexical representations for optimal search, (6) phonological recoding of word sequences in sentences, (7) ambiguity caused by errors in the preliminary phonetic representation, and (8) interpretation of prosodic cues to lexical identity. Previous models of speech perception, such as the motor theory, analysis by synthesis, and the Logogen, have not detailed solutions to all eight problems. The LAFS (*L*exical *A*ccess *F*rom *S*pectra) model is proposed here as a response to those issues; it combines expected phonological and acoustic-phonetic properties of English word sequences into a simple spectral-sequence decoding network structure. Phonetic segments and phonological rules play an important role in network compilation, but not in the direct analysis of the speech waveform during lexical search. There is no feature-detector stage in LAFS either. If viewed as a perceptual model, LAFS constitutes a simple "null hypothesis" against which to compare and refine alternative theories of acoustic analysis and lexical search.

Introduction

Recent spectrogram-reading experiments (Cole, Rudnicky, Reddy & Zue, 1978) have shown that the acoustic signal is rich in phonetic information. Without knowing anything about the words that are present, an expert spectrogram reader can produce a broad phonetic transcription that agrees with a panel of phoneticians from 80 to 90% of the time, depending on the scoring method used. Furthermore, perceptual experiments by Liberman & Nakatani (pers. comm.) indicate that listeners can transcribe nonsense names embedded in sentences (and obeying the phonological constraints of English) with better than 90% phonemic accuracy.

These experiments call into question the view that the speech signal is so impoverished of phonetic information that speech perception usually proceeds "top-down" with syntactic and semantic knowledge sources hypothesizing lexical candidates to be compared with aspects of the acoustic signal for verification. Of course there are listening conditions where noise or distortions force the listener to rely more heavily on expectations and higher-level knowledge to hypothesize words, but I believe that a bottom-up method of lexical access is an essential part of the normal speech decoding process. This paper

0095–4470/79/030279+34 $02.00/0

will be concerned with the process of lexical hypothesis formation from acoustic data. Little will be said about how such a bottom-up component of the speech understanding process interfaces with other components of a complete model of sentence perception.

There have been several recent efforts to build computer-based speech understanding systems that accept spoken input sentences within some limited domain, and respond with the correct answer better than 95% of the time (see Klatt, 1977 for a review). Of particular interest are the HARPY system (Lowerre & Reddy, 1978), which represents a large but finite set of sentences by a network of expected spectra, and the HWIM system (Klovstad, 1978; Wolf & Woods, 1978), which takes into account the phonological recoding of words and word sequences in normally spoken sentences. Examination of these systems has changed my views about how speech is normally perceived. Perhaps it is not wise to draw conclusions about the functioning of the human brain from analogies to computer algorithms, but the theoretical advantages of combining some of these strategies into a perceptual model are compelling.

A typical three-step machine method of lexical access is shown in Fig. 1(a). Parameters are extracted from the speech waveform, a phonetic transcription is derived, and then lexical hypotheses are proposed. Parameters might include formant frequencies (Zue & Schwartz, 1978), articulatory configurations (Wakita & Kasuya, 1977) or spectra (Lowerre & Reddy, 1978). The phonetic representation might be a distinctive feature matrix (Medress, 1969) or a lattice of segmental alternatives (Wolf & Woods, 1978). Lexical search might proceed in an analysis-by-synthesis mode at the syllable level (Weeks, 1974) or by precompiling phonological knowledge into a network of expected phonetic sequences for words, using scoring penalties for incorrect, missing, or extra segments in the input (Klovstad, 1978). The relative advantages among these choices are discussed in Klatt (1978*a*).

This chapter presents an alternative method of lexical access from acoustic input. In the next section, eight problems associated with word identification in running speech are

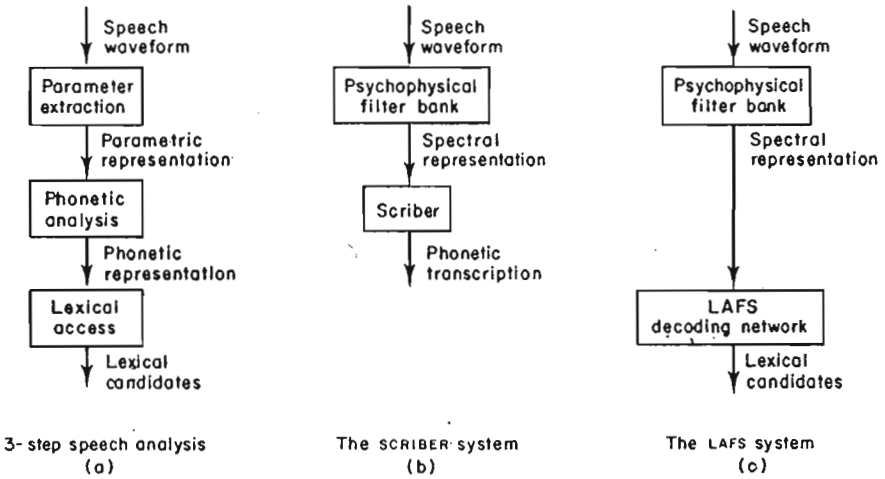

Figure 1 The block diagram of part (a) describes typical machine analysis procedures for bottom-up lexical access. Parts (b) and (c) outline the structure of two models of the early stages of speech perception to be described in the Sections "SCRIBER: a proposed solution to automatic phonetic analysis" and "LAFS: a proposed solution to the problem of lexical access".

identified. In later sections, two new computer systems, SCRIBER and LAFS, are proposed as potential engineering solutions to these problems. The LAFS system is then modified to form the perceptual model described in the section entitled "Implication for models of speech perception".

The SCRIBER phonetic transcription system described in detail later and shown here in Fig. 1(b) is proposed as an alternative to the more traditional methods of phonetic analysis. Knowledge of auditory psychophysics (such as critical bands, loudness, forward and backward masking, etc.) is used to derive an appropriate spectral representation for speech. Phonetic decoding rules then take the form of a network of expected sequences of static spectra for each possible transition between phonetic segments.

The LAFS system shown in Fig. 1(c) is proposed as a method for generating lexical hypotheses directly from a spectral representation of speech without first recognizing phonetic segments. Acoustic-phonetic knowledge and word-boundary phonology are precompiled into a decoding network of expected spectral sequences for all possible word sequences from the lexicon. This system avoids making possibly errorful early phonetic decisions and thus avoids the problems inherent in using an errorful phonetic transcription to search the lexicon.

The section entitled "Implications for models of speech perception" is concerned with modeling how humans generate lexical hypotheses from acoustic information. A new perceptual model of bottom-up lexical access is described that incorporates both SCRIBER and LAFS as components. The model departs from most current views of how speech is perceived in that phonetic segments and phonological rules play a role only in LAFS network compilation, and not in the direct analysis of the speech waveform during lexical search.

In the final section, the perceptual model proposed in the section entitled "Implications for models of speech perception" is compared with other models. One of these models, analysis by synthesis at the lexical level, is described in some detail in order to establish the relative advantages of precompilation of acoustic-phonetic and phonological relations to active synthesis of the same knowledge.

The problem

Before describing SCRIBER and LAFS, eight problem areas are identified that have plagued designers of speech recognition and speech understanding systems for decades. All have to do with the identification of words in spoken sentences, given some representation of the input acoustic waveform. The problems listed in Table I are endemic to speech com-

Table I Eight problem areas that must be dealt with by any model of bottom-up lexical access

1. Acoustic-phonetic non-invariance.
2. Segmentation of the signal into phonetic units.
3. Time normalization.
4. Talker normalization.
5. Lexical representations for optimal search.
6. Phonological recoding of words in sentences.
7. Dealing with errors in the initial phonetic representation during lexical matching.
9. Interpretation of prosodic cues to lexical items and sentence structure.

munication; they must be overcome by any speech processing system or model of human speech perception.

Acoustic-phonetic non-invariance

The acoustic manifestations of a phonetic segment are known to vary in different phonetic environments (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Klatt, 1978*b*) While many acoustic cues are largely context independent and thus form invariant properties of a given phonetic segment (Blumstein, Stevens & Nigro, 1977; Cole & Scott, 1974; Fant, 1974; Stevens, 1975) there are also context-dependent cues to the same phonetic distinctions. Listeners seem to be able to process these latter cues when the invariant cues are artificially removed (Delattre, Liberman & Cooper, 1955), and are probably able to make use of context-dependent information under normal listening conditions as well. What decoding strategy permits the listener to interpret cues that depend on phonetic context (so as to make optimum use of all the information contained in the speech waveform), especially when phonetic environment itself is not known with any certainty? More importantly, what are the perceptually relevant acoustic cues to each phonetic contrast, and how are cues combined?

Segmentation into phonetic units

Is segmentation an independent process, preceding phonetic labeling, or is it simply an automatic consequence of the phonetic decision process itself? If the speech waveform could first be segmented reliably into chunks corresponding to phonetic segments, the job of identifying each segment would be much simplified. Unfortunately, segmentation criteria depend on detailed knowledge of articulatory-acoustic relations, the permitted phonetic categories of the language in question, and other phonological constraints (see Klatt, 1978*a*; Cole *et al.*, 1978 for examples). Ambiguities in segment boundary locations are produced when the various articulators move asynchronously, as is often the case. A particular difficulty with feature-detector models of phonetic perception, such as the model described by Pisoni & Sawusch (1975) is the problem of interpreting detector outputs in order to align the columns of the derived feature matrix and see how many segments are present.

No matter how segmentation is accomplished, it appears that errors are inevitable. When a system commits itself to an error in an early stage of the analysis, such as a segmentation error, it is difficult for other components to recover and find the correct analysis. If a way could be found to defer or avoid segmentation decisions, overall system performance might improve.

Time normalization

Segmental durations are influenced by many factors, including speaking rate, locations of syntactic boundaries, syllable stress, and features of adjacent segments (Klatt, 1976*b*; 1979; Lehiste, 1970). A segment can vary in duration by a factor of two or three depending on its environment in the utterance. This kind of temporal variation clearly rules out the use of spectral prototypes of fixed duration in phonetic recognition. Some sort of time warping, time normalization, or method of ignoring the time dimension must be devised.

In addition, there are cases where the duration of an acoustic event can play a decisive role in a phonetic contrast of English (e.g. /ɛ/–/æ/ or /s/–/z/), but the durational dividing line between the two phonetic categories is sensitive to the factors listed above. Thus there are really two parts to the time-normalization problem: (1) how to ignore irrelevant

variations in segmental durations, and (2) how to incorporate durational information in selected segmental decision strategies when durational perturbations due to syntax, semantics, and stress are not known at this level. A phonetic transcription system can be designed to solve the first half of the time-normalization problem, but can never be entirely successful at the interpretation of durational cues to segmental contrasts. We are faced with the classical chicken-or-egg problem; phonetic decisions depend in part on lexical and syntactic factors that cannot be resolved until the phonetic decisions have already been made. This appears to be another example where the principle of delayed commitment is applicable: if possible, one should not make phonetic decisions prior to lexical hypothesis formation. The decoding of durational cues to syntactic structure and semantic emphasis is discussed more fully below under problem eight—interpretation of prosodic cues.

Talker normalization

Talkers differ in the length and general shape of their vocal tracts, in the articulatory-acoustic targets they use for each phonetic segment type, in their coarticulatory strategies as a function of stress and speaking rate, and in the dialect they employ (Stevens, 1972*c*). In addition, speech is heard in many different kinds of noise, reverberation conditions, and telephone channels. The variability created by these various factors does not seem to cause great difficulties for the listener, but little is known about the perceptual strategies used to normalize for different talkers and listening environments.

Lexical representations for optimal search

It is often tacitly assumed that the lexical representations used for matching during speech perception come from the same stored lexicon that is used for speech production (Liberman & Studdert-Kennedy, 1978), and that this single lexicon contains fairly abstract forms for morphemes (forms of the kind discussed by Chomsky & Halle, 1968). However, it seems clear that speech analysis routines cannot be expected to derive these abstract phonemic forms using bottom-up analysis (reasons are suggested by the examples of Table II below). The actual lexical representations used during matching are probably more nearly phonetic or acoustic in nature. These forms are either derived on the fly by generative rules, or (as is more likely given the computational efficiency) precompiled into some optimal form for rapid lexical search. The precise nature of the lexical representations used in bottom-up speech analysis is not known.

Phonological recoding

The expected phonetic realization of a word depends on the sentence context in which it appears (Oshika, Zue, Weeks, Neu & Aurbach, 1975). Consider for example the phonetic string observed for the spoken utterance "Would you hit it to Tom?" shown in Table II. No word boundaries are indicated in the phonetic transcription because acoustic cues to word boundary locations are rarely present within phrases (although the utilization of separate prevocalic and postvocalic allophones of liquids and voiceless plosives helps to constrain the possible locations of some word boundaries). Each of the simplifications listed in Table II can be described by general phonological rules. During speech production, such rules are assumed to operate on an underlying abstract phonemic representation for each word or morpheme. For example, an (optional) word-boundary phonological rule /d#y/ → [j] transforms the word-final phoneme /d/ and the word-initial /y/ into the phonetic segment [j] in "would you".

Table II Examples of word-boundary phonology

"Would you hit it to Tom"

[W ʊ J ə h I ɾ I t ə̥ t ɑ m]

1. Palatalization of /d/ before /y/.
2. Reduction of unstressed /u/ to schwa in "you".
3. Flapping of intervocalic [t] in "hit it"
4. Reduction to schwa and devoicing of /u/ in "to".
5. Reduction of geminate [t] in "it to".

In most models of lexical access, such modifications must be viewed as a kind of noise that makes it more difficult to hypothesize lexical candidates given an input phonetic transcription. To see that this must be the case, note that (a) pronunciation variants cannot be stored in the dictionary, since one doesn't want to accept [Jə] for "you" in the word sequence "are you", and (b) each phonological rule example of Table II results in irreversible ambiguity—the [J] observed in the sample phonetic transcription of Table II could be the first or last segment of a word like "judge", or it could be the surface manifestation of an underlying /d/≠/y/. The number of phonological phenomena is quite large and their effects on unstressed syllables can be dramatic, as suggested by the examples in Table II. Phonological recoding, both within words and across word boundaries, must be accounted for in a perceptual strategy. The significant amount of ambiguity introduced by cross-word-boundary phonological rules seems to support a "top-down" analysis-by-synthesis model of lexical access unless knowledge of the effects of these rules can be precompiled into an appropriate "bottom-up" decoding structure.

Dealing with phonetic errors

Even an ideal phonetic transcription component will make errors in the presence of environmental noises, talker variability, and other factors. Thus the lexical matching component must be able to find the (hopefully correct) best-scoring word even when no words match the input perfectly. The derivation of scoring algorithms for segmental substitution, omission, and insertion errors is difficult because some phonetic confusions are likely only in particular phonetic and stress environments. Very little is known about perceptually motivated scoring algorithms and decision strategies appropriate for lexical search.

Interpretation of prosodic cues

Prosodic cues (fundamental frequency contour, pattern of segmental durations, and intensity contour) are used by the talker to distinguish between stressed and unstressed syllables, to delimit syntactic units, to indicate contrastive stress or emphasis, and to signal psychological state or attitude toward the utterance (Klatt, 1976*b*; Lea, 1973; Lehiste, 1970; Lieberman, 1967). For purposes of lexical access, cues to the stress pattern can be quite useful. Many lexical alternatives can be ruled out if they do not have the right pattern of stressed, unstressed and reduced syllables.

Unfortunately, perturbations to prosodic contours that depend on syntactic, semantic, and psychological variables confound the situation and make interpretation of the stress pattern difficult. In addition, interpretation of syllable stress from prosodic variables is complicated by segmental factors. Some phonetic segments are inherently more intense, or they are of greater duration, or they perturb the fundamental frequency contour. The

listener appears to make stress judgments that are relative to these intrinsic properties of segments. As noted earlier, segmental decisions such as /ɛ/–/æ/ depend on duration cues that can only be interpreted with certainty after the stress pattern is known; it seems that segmental judgments and stress judgments must be computed simultaneously and interactively.

Of the eight problems outlined above and in Table I, the first four are addressed by the SCRIBER phonetic transcription system. The second computer system to be described, the LAFS lexical access algorithm not only takes advantage of the solutions embodied in SCRIBER, but also adds strategies that effectively deal with the final four problems. Sections describing the computational algorithms of SCRIBER and LAFS are followed by a discussion of the relations between these components and models of speech perception.

SCRIBER: a proposed solution to automatic phonetic analysis

This section is concerned with the specification of a new computer algorithm for generating a phonetic transcription of the acoustic waveform corresponding to an unknown English sentence. The system is called SCRIBER, and is presently under development in the Speech Communication Laboratory at M.I.T. This preliminary report is concerned only with the design philosophy of the system since there are no results to report as yet.

A tentative set of about 55 output phonetic categories has been selected. The inventory of phonetic segment types is large enough to preserve distinctions useful in lexical decoding (e.g. postvocalic allophones of the liquids, unstressed and unreleased allophones of the plosives, etc.), but it is by no means intended to represent a narrow phonetic transcription.

Representation of acoustic-phonetic knowledge

As an engineering approximation, it is assumed that transitions between phonetic segments can be represented succinctly and accurately by sequences of a few static spectra. For example, Fig. 2 illustrates a sequence of four spectra used to characterize the transition between [t] and [a] in the phrase "the top of the hill". Such a transition from the middle of one phone to the midpoint of the next is called a diphone. It has been argued that the coarticulatory influences of one phone on its neighbors do not usually extend much further than half-way into the adjacent phones (Peterson, Wang & Sivertsen, 1958; Gay, 1977). To the extent that this approximation is true (see Lehiste & Shockey, 1972 for supporting perceptual evidence), diphone concatenation captures much of the context-dependent acoustic encoding of phonetic segments.

There exist a number of special cases that require attention in a diphone system. For example, a vowel followed by a nasal can be nasalized to a variable degree. The SCRIBER system is designed to produce, as output, the intended non-nasalized vowel. The technique employed is to define two (or more if necessary) alternative spectral sequences that describe the same diphone—one with a nasalized vowel, and one without. Similar solutions are required to decode other optional coarticulatory phenomena and to deal with certain unstressed allophones.

The choice of the diphone as the unit used to relate acoustic and phonetic levels is not central to any of the models to be described. A diphone dictionary is a convenient tabular way of cataloging acoustic-phonetic relations, but the same relations could, in principle, be described by rules (if the appropriate rules were known) or in terms of a dictionary describing the spectral manifestations of larger units such as triphones (Wickelgren, 1969), demisyllables (Fujimura & Lovins, 1978), or syllables (Studdert-Kennedy, 1976).

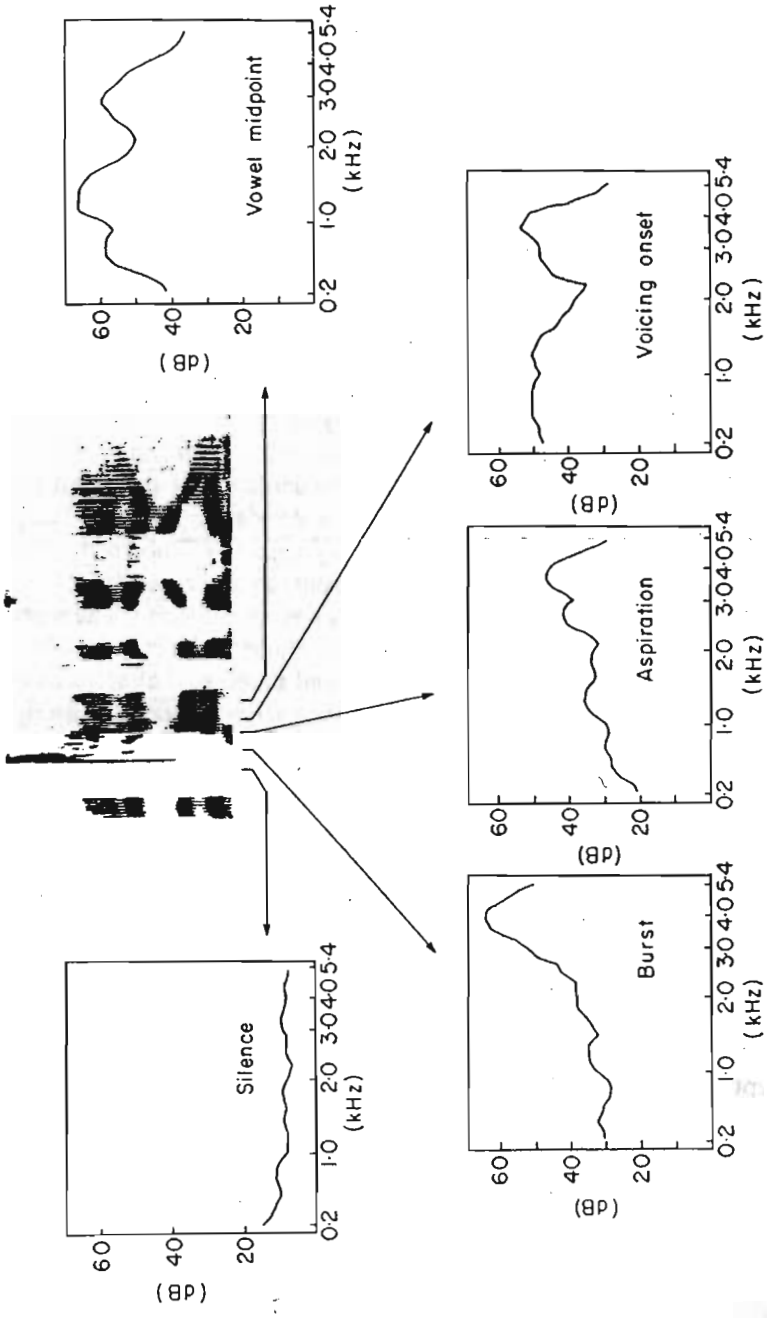

Figure 2.

A broadband spectrogram is shown of the phrase "the top of the hill" in order to indicate times at which spectra were computed so as to characterize the transition from the middle of closure for [t] to the middle of the vowel [a] in the SCRIBER phonetic decoding network.

The spectral representation that has been chosen is based on the psychophysical considerations given in Table III (Klatt, 1978*a*). A short-term spectrum is computed every 10 ms using a set of 30 overlapping critical-band filters. Several examples of these spectra are shown in Fig. 2. It is up to the experimenter to select sufficient sample spectra to characterize each possible phonetic transition of English. There are 55 phonetic segment types in the inventory of SCRIBER, but many of the 55-by-55 possible acoustic transitions do not occur. Only about 2000 *diphones* are phonologically permissible in English. Each of these diphones is thus characterized by a sequence of three or four spectral templates, as in Fig. 2.

Table III Psychophysical consideration in the design of a spectral representation for speech processing

1. Include frequency components from at least 270 to 5600 Hz since this is the minimum passband for which there is no measurable loss in intelligibility when compared with systems containing wider bandwidths (French & Steinberg, 1947).
2. Include a dynamic range of at least 50 dB so as to adequately represent spectra of both the intense and weak speech sounds.
3. Provide a temporal resolution of about 10 ms since this is the best current guess as to the shortest spectral window employed by the auditory system, and since otherwise certain rapid formant transitions and brief plosive bursts might be missed.
4. Take into account the observation that our ears cannot resolve individual harmonics of a voiced sound if the harmonics are spaced within a critical bandwidth of about a quarter of an octave (Houtgast, 1974; Plomp & Mimpen, 1969; Sharf, 1970).
5. Take account of the fact that the contribution to intelligibility from different portions of the spectrum is not uniform (French & Steinberg, 1947). The relative importance to speech intelligibility of different frequency components is in good agreement with a theory stating that each critical bandwidth contributes about equally to intelligibility, at least over the range from 270 to 5600 Hz.
6. Design the slopes of the critical band filters so as to account for the spread of masking (i.e. low frequencies mask weak higher-frequency components better than vice-versa, so the filters have more gradual low-frequency skirts).
7. Express the output of each filter in dB (because decibels are an approximately equal-interval scale for loudness), and quantize filter outputs to about 1 dB (because the just-noticeable difference for changes to formant amplitudes change is 1 dB or more, depending on the circumstances (Flanagan, 1957)).
8. Process only the *magnitude* of the spectrum because the phase spectrum is too unpredictable to be used in phonetic decoding.
9. Use a number of overlapping critical-bandwidth filters sufficient to discriminate spectral changes caused by formant frequency changes of about 3 to 5% since this is the just-noticeable difference for a formant frequency shift (Flanagan, 1957).
10. Employ a pre-emphasis filter based on a pure tone threshold curve which indicates that there is an effective emphasis of frequencies in the 2 to 3 kHz range. Use equal-loudness contours to compute the growth in loudness with increases in signal intensity (Zwicker, Terhardt & Paulus, 1979).

There can be template sharing for portions of diphones that are acoustically similar. For example, Fig. 3 indicates expected spectral sequences for the prestressed aspirated consonant [t] followed by any stressed vowel of English. The decoding structure summarizes the obvious fact that the closure (silence) spectrum for [t] is the same before any vowel, and the observation that onset spectra for [t] are virtually identical before all front vowels, identical before all back unrounded vowels, and identical before all rounded vowels (Klatt, 1978*b*; Zue, 1976). As indicated in the figure, spectral characteristics observed during aspiration are dependent on both [t] and the vowel, since formant onset values depend on the vowel (Klatt, 1978*b*). In general, a new spectral template is defined for each distinctive spectrum that is observed in a phonetic transition. If the transition is

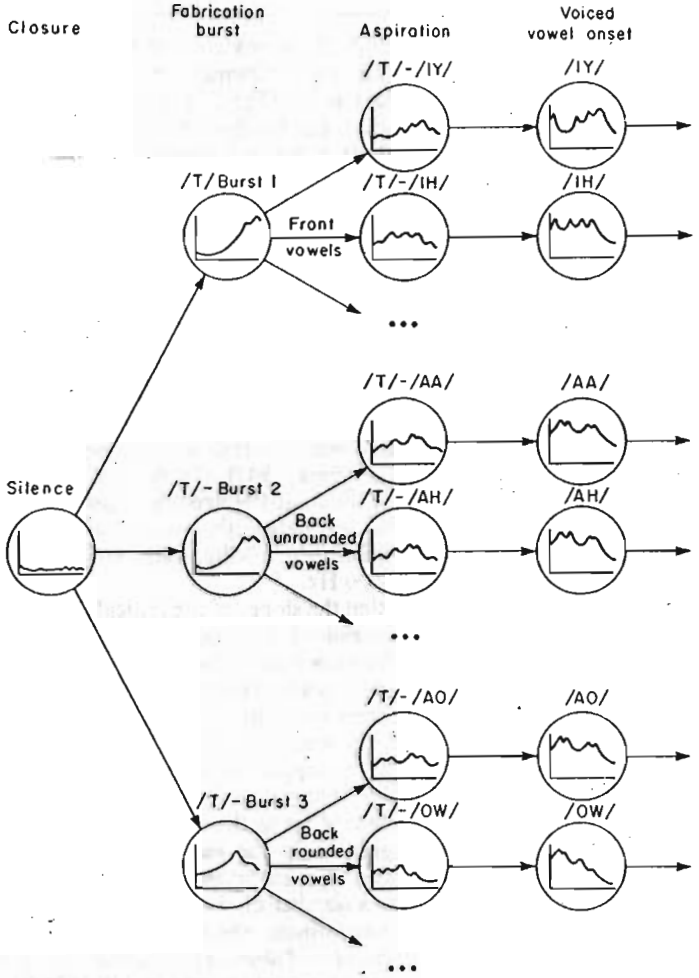

Figure 3. A small portion of the SCRIBER spectral sequence phonetic decoding network is shown to illustrate the defining characteristics of prestressed prevocalic [t]. Each state of the network (circle) is characterized by a spectral template (dB versus frequency, as shown inside the circle). Not shown are durational constraints in the form of a feedback path to each state indicating the expected number of input 10-ms spectra that can be associated with each state during recognition.

rapid, templates may be defined as often as every 10 or 20 ms, while for gradual spectral changes, few templates are defined per unit time.

A complete spectral sequence phonetic decoding network is compiled automatically from transition definitions of the type shown in Fig. 2. The network is highly interconnected since all possible phonetic transitions must be represented by spectral sequences. However, there exists a relatively simple recognition strategy for utilizing this compact representation of acoustic-phonetic knowledge.

Recognition strategy

Once a network has been created, the input waveform is analyzed by computing a spectrum every 10 ms (using overlapping 25·6 ms chunks of windowed waveform). The recognition strategy consists of comparing this input spectral sequence with spectral templates of the network. The idea is to find the path through the network that best represents the observed input spectra. This path defines the optimal phonetic transcription.

A Euclidean distance metric involving differences in dB in selected frequency bands will be used to compute matching scores for each input spectrum with the spectral templates, in a manner very similar to the strategy employed in the HARPY sentence recognition system (Lowerre & Reddy, 1978). The metric has some perceptual validity for static spectral comparisons (Lindblom, 1978), but it will no doubt have to be modified to take into account changes in average speech spectrum and noise background associated with a new speaker or a new recording environment (Klatt, 1976*a*). Modifications may also be needed for comparison of template sequences so as to emphasize dynamic over static properties of the speech signal. For example, a long vowel probably should not contribute much more to dynamic distance than a brief plosive burst.

Input spectra are processed one at a time and used to try to extend the most promising "phonetic hypotheses". A hypothesis consists of (a) a pointer to a state in the decoding network, (b) the duration of the input that has been associated with that state, (c) a cumulative spectral matching score, and (d) the last two phones of the implied phonetic transcription. An hypothesis is extended by assigning the new 10-ms input spectrum to the current state in the network or to one of the network states that can be reached from this state. An hypothesis will generate several daughter hypotheses if alternative paths from the current network state score reasonably well.

When the best-scoring hypothesis in the (ordered) list of all current alternatives reaches a place in the network calling for the output of a phonetic symbol, it is assumed that the earlier of the two previous phonetic symbols that have been saved with the hypotheses is correct. The phonetic symbol is outputed and all hypotheses not possessing this symbol are pruned from further consideration. A large number of best-scoring hypotheses are pursued in parallel during a HARPY-like "beam search" of all alternative paths having scores that are within some fixed distance of the best-scoring path to date (Lowerre & Reddy, 1978). Given this strategy, SCRIBER evaluates nearly all reasonable phonetic transcriptions of the input and selects the optimum transcription.

Phonetic non-invariance

How does SCRIBER deal with the first four problems listed in Table I? As a solution to the acoustic-phonetic non-invariance problem, SCRIBER incorporates diphone definitions. Any acoustic cue that is dependent on immediate phonetic context can be represented in a SCRIBER decoding network. Thus, for example, the burst spectrum for [t] is expected to have its most prominent spectral peak at lower frequencies before rounded vowels than

before unrounded vowels (as shown in Fig. 3), and systematic differences in formant motions following [t] release into various vowels are described by individual aspiration templates for each [t]-vowel diphone. To the extent that acoustic invariance is present, states in the network are combined (for example, the [t] burst is represented by the same template before all front vowels in Fig. 3).

Diphone prototypes have been defined in terms of sequences of spectra in order to test the simplest possible hypothesis concerning which acoustic cues are most salient to each phonetic decision. It is hoped that, if the right metric can be devised for comparing spectra, SCRIBER will perform well and there will be no need to postulate a representational level between spectra and phonetic segments. The alternative is to interpose certain kinds of property detectors or phonetic feature detectors that extract particular attributes from the spectra. Such a level will not be included in SCRIBER unless the simpler model fails to account for various natural acoustic-phonetic distinctions (see Stevens, 1972*a* for a partial list). Thus while sequences of spectra are not acoustic cues in the usual sense, they imply a theory in which spectra form a Gestalt or holistic unanalyzed representation rather than the input to a system of feature analyzers.

Segmentation

There is no explicit segmentation step in the SCRIBER decoding strategy. In a sense, all possible segmentations (alternative assignments of input spectra to network states) are evaluated in parallel. Since the input is not segmented before phonetic labeling, there is no need to develop strategies for correcting errors in segmentation. The final transcription provides an implicit segmentation since each 10-ms input spectrum has been associated with a particular state of the best-scoring path through the network. Therefore, durations of acoustic events can be computed if relevant to a phonetic contrast (see next paragraph).

Time normalization

The sequence-of-spectra concept is attractive for a number of reasons. For example, if desired, one could allow acoustic events to have any arbitrary duration without penalty, and irrelevant durational variability would be ignored. However, it is well known that duration is important for a number of phonetic contrasts, and some mechanism for incorporating durational constraints in the network representation is essential. To achieve this goal, the system is augmented in the following way. The expected duration of the input to be associated with each spectral template of a phonetic transition definition can be added to the diphone definitions for those cases where duration is deemed important. The result is that any state in the network of Fig. 3 can be assigned an explicit feedback path specifying the expected number of input spectra that can be associated with that state during recognition. For example, the number of input spectra associated with the burst spectrum plus the number associated with the aspiration spectrum in Fig. 3 should be about 5 (50 ms, i.e. the voice onset time should exceed about 25 ms) for a prestressed [t] to be recognized. In those cases where duration is determined to be important to a phonetic contrast, differences between expected and observed durations of the input assigned to a spectral template contribute to the distance score for a hypothesis. In this way, durations of acoustic events can be measured and compared with an accuracy that seems consistent with the relatively large (25 ms or more) durational just-noticeable differences observed during sentence perception (Klatt & Cooper, 1975).

Rate of spectral change is not represented by this means because it appears that cases where rate seems important [e.g. in distinguishing between /ba/ and /wa/, (Liberman,

Delattre, Gerstman & Cooper, 1956] depend more on the duration of the initial [w]-like spectrum, and thus can be better represented by specifying the expected duration of an initial steady state spectrum, and specifying that a certain spectral sequence be traversed. Rate of formant transitions or rates of other spectral changes are difficult to represent in discrete networks of this sort. One cannot easily constrain the relative duration of each component template of a transition definition—only the overall duration of the transition and/or the duration of any initial or final steady states. If rate turns out to be a perceptually important *independent* acoustic cue, this would constitute evidence against the template-sequence approach outlined here.

As argued earlier, variations in segmental durations due to speaking rate, syntactic factors, stress, and phonetic environment make it very difficult to rely on absolute durational constraints to distinguish among phonetic segments. The SCRIBER system can be set up to ignore irrelevant variations in segmental duration, but higher-level variables that influence segmental durations contribute durational ambiguity that simply cannot be overcome. Duration ratios among adjacent acoustic events may serve as useful speaking-rate-invariant cues for certain phonetic contrasts (Port, 1978), but a minimum-use-of-duration strategy still seems wise in any attempt to build a phonetic recognizer. The inability to make effective use of durational (and FO) cues to segmental contrasts is one of the primary reasons why I feel that a phonetic transcription component is not an appropriate driver for lexical search.

Talker normalization

One criticism of previous template models of speech recognition is that they cannot be modified very easily to handle different talkers. There is considerable variation in the details of spectra characterizing phonetic segments spoken by men, women, and children. On the other hand, acoustic patterns observed for adult talkers are more similar in a critical-band spectral representation than one might expect (Searle, Jacobson & Rayment, 1979). This observation lends support to a talker normalization procedure proposed by Lowerre (1977). He restricted the HARPY sentence recognition network to contain only 98 different spectral template types, and all sentences had to be represented in terms of sequences of spectra drawn from these 98 templates. Templates were modified incrementally toward spectra seen for a new speaker in the following way. If a sentence could be recognized using templates representative of an "average talker" (the sentence error rate was about four times as great as when talker-specific templates are available), then the observed input spectra were used to modify those spectral templates of the network that were matched during recognition.

An added advantage of this approach is that it captures some idiosyncratic aspects of acoustic targets employed by each talker in realizing different phonetic segments. For example, if the speaker habitually uses a fronted /u/, the appropriate template(s) converge toward spectra that reflect this habit. The network is intended to represent the acoustic-phonetic characteristics of a particular dialect of English, while the spectral templates represent acoustic targets that are talker-dependent. The separation of the SCRIBER system into talker-dependent templates and a talker-independent knowledge network has important theoretical implications. Speech processing by man and machine would be considerably simplified if such a separation could be experimentally validated.

In the SCRIBER system, more than 98 spectral templates will be required to make fine phonetic contrasts (about 300 may be sufficient), but the dynamic talker normalization procedure of HARPY can still be applied. In addition, several generalized methods of talker

normalization will be investigated, such as starting with an average female template if the new talker seems female, or modifying all templates on the basis of average spectral properties of a new voice, or estimating vocal tract length, or saving template sets for familiar voices.

When compiled into a decoding network, SCRIBER is not particularly large. A complete 2000-diphone inventory requires an average of about two new states per diphone, resulting in a network of about 4000 states and 6000 paths. (This is substantially smaller than the 15000-state HARPY sentence-decoding network that can recognize ten-to-the-eight different sentences.) Sentence decoding then involves a large number of similar computations that can be performed in real time on a present-day fast digital processor such as the Floating Point Systems AP-120B.

Advantages of SCRIBER

The main advantages of SCRIBER are (1) the possibility of embedding all acoustic-phonetic knowledge concerning English (including phonological constraints on permitted phonetic sequences) into a single uniform network representation, (2) the ability to produce a phonetic transcription by simultaneously evaluating the scores for most of the likely alternative phonetic transcriptions, and (3) no need for explicit phonetic segmentation. Knowledge appears in a transparent form (the dictionary of spectral sequences for each phonetic transition) that makes optimization relatively easy. If it is successful, SCRIBER has possible applications as a limited-performance phonetic typewriter, as a "front end" for a computerized speech understanding system, as an aid for the deaf, and as a part of a model of speech perception (discussed later).

LAFS: a proposed solution to the problem of lexical access

The LAFS (lexical access from spectra) system is a computer algorithm for efficient accurate lexical search. LAFS avoids explicit phonetic transcription by precompiling knowledge of acoustic-phonetic relations into lexical definitions, in a way that is based on SCRIBER. The system deals with ambiguity generated by phonological recoding rules by precompiling knowledge of the rules into a decoding network.

Lexical representations

The first step in the design of LAFS is to construct a tree of expected *phonemic* sequences for all words of the lexicon, as shown in Fig. 4(a). An abstract phonemic lexicon is assumed as a starting point because of the many theoretical advantages of postulating abstract underlying forms for words and morphemes (Chomsky & Halle, 1968), even though the psychological lexicon may not include some of the more abstract, less productive rules (Ohala, 1974). The phonemic lexicon is organized into the form of a tree [Fig. 4(a)], such that words having the same initial phoneme sequence share nodes (phonemes) and branches until the words diverge in phonemic representation. Initial portions of words are shared so as to save storage, increase search speed, and facilitate application of phonological rules. Of course, a pair of words cannot share tree nodes if the words react differently to phonological rules due to stress differences or other factors.

Precompiled phonological rules

Phonological rules are used to derive phonetic forms for each word. Rule application often depends on characteristics of adjacent words, so the lexical tree is first modified in the following way. The end of each word in the tree is attached to all word-beginning

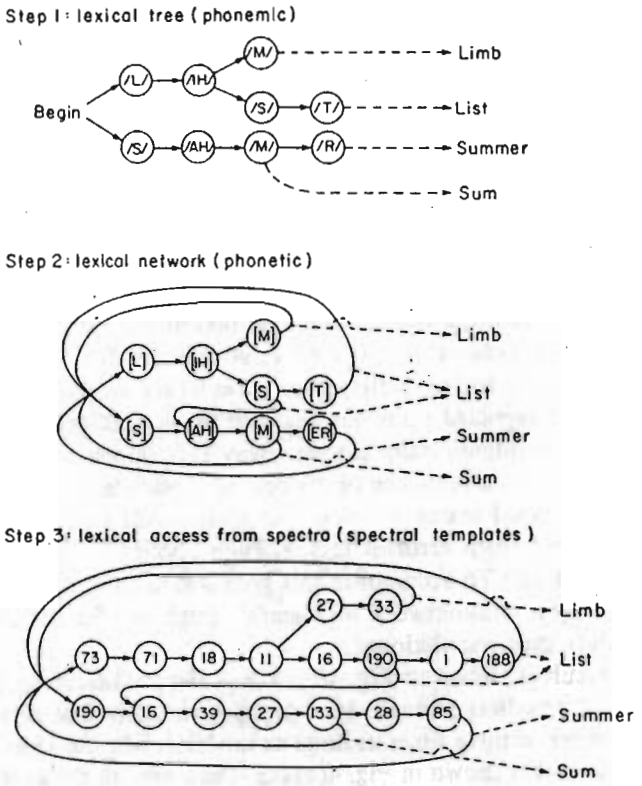

Figure 4. The LAFS lexical-access-from-spectra decoding network that is shown in part (c) is derived by first constructing a tree of phonemic representations for all words in the lexicon, a portion of which is shown in part (a), and then connecting all word terminations to word beginnings and applying a set of phonological rules in order to form a phonetic sequence lexical decoding network, as shown in part (b). The numbers inside the states in the spectral sequence decoding network of part (c) refer to spectral templates from an inventory of about 300.

states. Then a set of phonological rules are applied to replace each phoneme by an appropriate phonetic allophone, delete or replace some segments, and modify the connectivity pattern (Klovstad, 1978). The result is a phonetic-sequence lexical decoding network of the type shown in Fig. 4(b), representing expected phonetic properties of all possible (grammatical and ungrammatical) word sequences of the lexicon. Cross-word-boundary phonological phenomena that must be described include the possible insertion of a silence or juncture phone such as the glottal stop, normal phonetic coarticulation, and various simplifications such as palatalization, flapping, [t]-deletion, etc. (Oshika *et al.*, 1975).

For example, the effect of the [st ≠ s] → [s] phonological rule in Fig. 4(b) is to create an extra path from near the end of words ending in [st] to the second phonetic segment of words beginning with [s], so that a word pair such as "list some" can be recognized when spoken in the normal way, i.e. without the [t]. Since the rule is optional, the network must represent both alternatives, and the path from the final [t] of "list" to words beginning with [s] is therefore not broken. In the past few years, phonologists have developed formal rules of considerable predictive power (Chomsky & Halle, 1968; Cohen & Mercer, 1975;

Oshika, *et al.* 1975; Woods & Zue, 1976) that should prove useful in the context of lexical access.

If lexical access is attempted from a phonetic transcription, even given this decoding network in which words are represented in terms of phonetic segments, and phonological rule phenomena are represented by the connectivity pattern, one still requires a sophisticated matching strategy to select words corresponding to the derived phonetic string. A metric is needed to determine penalties for mismatches and for segmental intrusions or deletions because the automatic phonetic analyzer will make many errors of these types. Experience with the BBN lexical decoding network has shown that metrics to handle errors are very important in that unexpected transcription errors may result in a fatal rejection of the correct word (Wolf & Woods, 1978).

The ideal way to deal with transcription errors would be to go back to the acoustic data to see if the expected phonetic sequence for a word scores reasonably well. A phonetic transcription intentionally throws away this information, reducing large amounts of acoustic data to a sequence of discrete phonetic elements, and thus, it has been argued, makes the lexical search problem computationally tractable. Why not avoid the problem of recovering from errorful intermediate phonetic decisions by not making phonetic decisions at all? To accomplish this goal within LAFS, each state transition in Fig 4(b) is replaced by a mini-network of spectral templates. Each min-network is obtained from the SCRIBER diphone dictionary.

The result is shown in Fig. 4(c), a lexical-access-from-spectra decoding network that has no intermediate phonetic level of representation. The network is quite large, but only on the order of three times as large as the lexical decoding network made up of phonetic segments that is shown in Fig. 4(b) (i.e. there are, on the average, about three new states required to represent a phonetic transition).

Recognition strategy

The LAFS network of Fig. 4(c) is very similar in structure to SCRIBER, and the decoding strategy of SCRIBER, i.e. find the best path using a simple spectral metric, can be applied. A lexical decision confirming the presence of word X is made when all alternative word sequence hypotheses not containing this word are unlikely to increase in score as more of the input waveform is processed. Hypotheses not containing the identified word are then pruned from further consideration in order to minimize the number of alternative word sequences that need be considered in parallel.

In a system containing a large lexicon, one cannot postpone a lexical decision very long because too many alternative partial phrases would then have to be considered. Hopefully, in most practical applications, decisions can be held in abeyance for up to 0·5 s after the end of the hypothesized word, and the lexical hypothesis buffer need not grow any larger than about 1000.

Interpretation of prosodic cues

Not only is it possible to evaluate all reasonable lexical alternatives simultaneously in LAFS, but one can specify tighter durational, voice-onset-time, and stress-related constraints on acoustic events associated with words in LAFS than in SCRIBER because one knows more about the expected stress pattern and phonetic environment. For example, a shorter vowel duration is expected in "better" than in "bet", although both contain the same stressed vowel; such a constraint is easily applied in LAFS, but cannot be handled by a phonetic decoder.

The LAFS system also offers a mechanism for preliminary hypotheses to be formed concerning the syntactic structure of the sentence and the locations of semantically important words. Since the LAFS network is derived from lexical representations, one can specify not only the expected phonemic string and the expected durational pattern, but also the expected fundamental frequency contour and the intensity envelope that would normally be seen in a stressed (or unstressed) sentence environment. Differences between expected and observed prosodic variables either signal that the lexical candidate is inconsistent with the input, or the differences might be interpretable as cues to syntactic structure. For example, an unusually long final syllable in a word would indicate the presence of a phrase or clause boundary. Other possibilities include identifying the first part of a compound by a shorter-than-usual word duration, or detecting an emphasized word by a higher-than-usual fundamental frequency peak. Assuming that the network contains absolute values of prosodic cues to be expected, some form of normalization will be required to compensate for average speaking level, speaking rate, and fundamental frequency range employed by the current talker before direct comparision of observed and expected data is possible.

While duration and intensity contour ought to be effectively interpreted at a lexical level, there is reason to question whether FO can be processed in this way. A word receives a number of alternative FO contours depending on sentence type (statement/question), syntactic position, and emotional nuances imparted by the speaker. Thus to interpret FO within LAFS, it may be necessary to have some current hypothesis in mind as to the prosodic interpretation of the previous input.

Morphemes versus words

If the lexicon is broken down into morphemes (e.g. books = book + s, baseball = base + ball), there can be considerable savings in both storage and processing time. Allen (1973) has assembled a morpheme dictionary that can represent at least ten times as many English words as there are morphemes. LAFS should probably be organized in terms of the more common morphemes, but for recognition purposes, a lexical decoding network must keep separate representations for morphemes that change pronunciation when bound together (e.g. applicability = apply + able + ity). Even so, an English lexicon containing e.g. 15000 morphemes would result in a phonemic tree of only about 50000 states, and the resulting LAFS decoding network would have less than 150000 states.

Relation to HARPY

The HARPY speech recognition system (Lowerre, 1976; Lowerre & Reddy, 1978) represented a finite set of sentences in terms of a network of words, represented words in terms of phonemic sequences, used a few phonological rules to select allophones and modify the connectivity pattern across word boundaries, and represented each of 98 phonetic segment types in terms of a single spectral template. While LAFS is based on these concepts, it differs from HARPY in several ways. For example, the set of acceptable input sentences is unbounded in LAFS. More importantly, LAFS has augmented abilities to characterize acoustic characteristics of phonetic transitions via diphone and triphone definitions. LAFS incorporates a better motivated spectral representation and distance metric than the HARPY linear prediction spectrum and minimum residual error metric. Finally, complex phonological recodings within words and across word boundaries can be expressed within LAFS, and prosodic cues can be interpreted.

Advantages of LAFS

LAFS has been designed to deal with all eight problems identified in Table I. The first four problems (acoustic-phonetic non-invariance, segmentation, temporal variability, and talker variability) are addressed by using the spectral-sequence diphone definitions and recognition strategy employed by SCRIBER. The fifth problem, how to represent words of the lexicon for optimal search, was solved by converting abstract phonemic forms into spectral sequences. The sixth problem, how to take into account phonological recoding across word boundaries, was solved by applying a set of phonological rules to augment and adjust the connectivity pattern of the lexical network. The seventh problem, how to recover from errorful phonetic decisions, has been nullified by not making intermediate phonetic decisions. Error recovery is still an issue in that background noises or mispronunciations can corrupt the input. Recovery from these distortions depends on the ability of the beam-search strategy to find the best path through the lexical network even when no words score very well. The eighth and final problem, interpretation of prosodic cues, has been solved by storing expected prosodic attributes of words in the lexical decoding network and by interpreting deviations from these expectations either (1) as an indication that the lexical hypothesis is not compatible with the input, or (2) as cues to syntactic and semantic structure.

Thus, in theory, a LAFS processor has the capability of representing all of the acoustic-phonetic and phonological knowledge needed to recognize words in spoken sentences. Lexical hypotheses can be generated rapidly and more accurately in a LAFS structure than in any two-step model (phonetic recognition followed by lexical access) containing the same acoustic-phonetic and phonological knowledge. Two-stage models violate the principle of delaying absolute decisions until all of the relevant information is available, and errors thereby introduced cannot always be overcome. LAFS will make fewer errors than SCRIBER for another reason: LAFS does not evaluate all phonologically possible phonetic sequence alternatives—only phonetic sequences that make up English words—and the consideration of fewer alternatives means fewer chances to make an error. The cost of recasting LAFS as a two-stage model would be both a decrement in performance and a need to add strategies for comparing errorful phonetic strings with expected phonetic strings; these strategies are totally unnecessary in a model that does not make phonetic decisions.

Implications for models of speech perception

Do the computational strategies of SCRIBER and LAFS have any relation to plausible models of speech perception? I believe that it is worthwhile to seriously entertain this possibility. A perceptual model based on LAFS may turn out to be too simple-minded to stand the test of time, but it can serve as an excellent framework for asking new kinds of experimental questions.

Figure 5 presents one such model of speech perception in the form of a block diagram. The model consists primarily of a LAFS bottom-up lexical hypothesis component. Tentatively, it is also postulated that the model also include a SCRIBER phonetic transcription component that is used for adding new words to the LAFS decoding network, and perhaps also for early verification of top-down lexical hypotheses.

Normal mode of lexical access

An input speech waveform ([1] in Fig. 5) is transformed into a sequence of spectra [2] by a spectral analysis component analogous to the peripheral auditory system. The LAFS

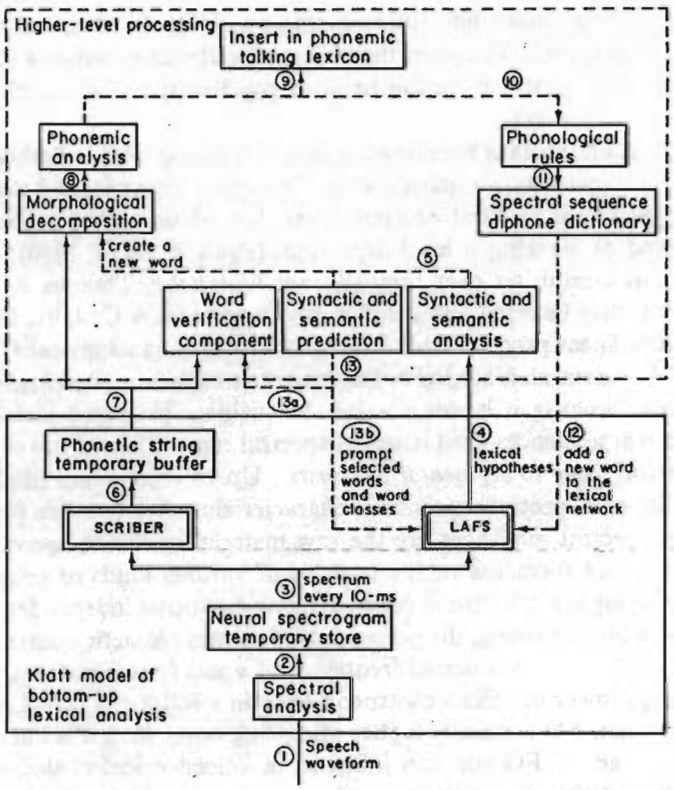

Figure 5. Block diagram of a proposed perceptual model of lexical access (enclosed in solid lines) and how it interfaces with higher-level components (enclosed in dashed lines). The numbers are included to aid the reader in following the explanation of information flow in the text.

lexical-access-from-spectra component reads information off the resulting "neural spectrogram" temporary store (a name coined by Peter Bailey to draw an analogy with a broadband spectrogram) in a single left-to-right pass by processing the input in 10-ms sequential chunks. The box labelled "LAFS" in Fig. 5 includes the spectral sequence lexical decoding network and the buffer of active lexical hypotheses generated by the recognition strategy. As output, LAFS produces lexical hypotheses [3] that are analyzed by higher-level syntactic and semantic components [4]. The output of LAFS consists not only a sequence of best-scoring words, but also matching scores that are used by the higher-level components to decide whether to accept a bottom-up lexical hypotheses.

There is no reason to suppose that the peripheral auditory system transforms speech into a sequence of discrete spectra sampled every 10 ms, as in LAFS. However, both the spectral representation and the LAFS network could be reformulated as a quasi-continuous model by sampling the spectrum every 1 ms, or even more frequently. It appears that there would be little change in performance if this were done since speech waveforms do not change that rapidly, and spectra sampled using a 10-ms time weighting window cannot change very much when the window is shifted only 1 ms forward in time.

More important issues are the duration and shape of the window over which short-term spectra are computed. The 25.6 ms Kaiser window that is employed in LAFS has an effective duration of about 10 ms. This value was chosen so as to maximize the accuracy of a

spectral estimate while still retaining an ability to track the rapid spectral changes that occur in speech. To model the human, the choice of window (or windows) will ultimately have to be justified on the basis of psychophysical data, physiological data, and LAFS performance data.

A second issue concerning the neural spectrogram is whether the spectrum is computed in the same way no matter what the signal. Physiological and psychophysical evidence suggests that spectral computations depend on signal properties in several ways: the spread of masking is level dependent (Egan & Hake, 1950); the response to stationary signals diminishes over time (Kiang, Watanabe, Thomas & Clark, 1965); and sudden onsets may be represented differently (Leshowitz & Cudahy, 1975). There is also evidence of non-linear processes that have the effect of enhancing peaks in vowel spectra (Houtgast, 1974). Research is needed to improve the simulation of preliminary spectral analysis in the model because it is not possible to evaluate proposed metrics for the comparison of spectral sequences until issues of spectral representation are settled.

Other cues to segmental contrasts Up to this point, emphasis has been placed the utility of a spectral sequence in characterizing each possible phonetic transition. I believe that spectral sequences are the raw material on which speech perception strategies are based (not formants or the outputs of various kinds of property detectors), but some qualifying remarks are in order. At least two other independent dimensions are known to play a limited role in the perception of certain phonetic contrasts.

One is the fundamental frequency of vocal fold vibrations (FO). For example, FO is usually lower in voiced obstruents than in a following vowel, while following a voiceless obstruent, FO is usually higher at voicing onset than it is later in the vowel (Lea, 1973). This kind of FO cue can influence a voiced-voiceless decision (Haggard, Ambler & Callow, 1970). In addition, a strictly spectral explanation, such as the hypothesis that an FO increase changes the amount of low-frequency energy in the spectrum, cannot account for the perceptual influence of the FO contour (Massaro & Cohen, 1976).

A second acoustic dimension that can influence a phonetic decision, but which is not subsumed by the proposed spectral sequence prototype for a phonetic transition, is the degree to which the spectrum is periodic (as in a vowel), aperiodic (as in a voiceless consonant), or contains both low-frequency periodicity and high-frequency aperiodic noise (as in a voiced fricative or voiced /h/). Differentiating between, e.g. a voiced /h/ and a vowel on the basis of spectral cues alone can be difficult. Perceptual data on the importance of a "degree-of-periodicity" cue are not available. Nevertheless, a measure of the degree of periodicity in the spectrum above about 1 kHz might be a useful acoustic parameter. It appears that the auditory system would have no difficulty in computing such a periodicity measure at high frequencies (Searle *et al.*, 1979).

The role of formant frequencies No mention has been made of formant frequencies and formant motions as possible cues for phonetic perception. In the acoustic theory of speech production, formant frequencies play a central role, characterizing the natural resonant modes of the vocal tract for a given articulatory configuration (Fant, 1960). However, automatic extraction of formant frequency information from the speech waveform is a difficult engineering task. It is still tacitly assumed by many that formant frequencies are psychologically real dimensions employed in perceptual decoding strategies (Delattre *et al.*, 1955; Carlson, Fant & Granstrom, 1975). We have no perceptual data that would refute this assumption, but there are several reasons to question its plausibility. For example, occasional formant tracking errors should result in dramatic errors in phonetic perception, whereas observed phonetic errors demonstrate a strong tendency to be

acoustically similar to the intended vowels and consonants (see, e.g. Miller & Nicely, 1955). As we have argued, absolute decisions at any level below the word (parametric representation, phonetic feature representation, or segmental representation) should be avoided if at all possible for optimal lexical decoding.

Metrics for spectral comparisons

Assuming that a reasonable characterization of the information residing in the neural spectrogram can be established, simple metrics for determining the similarity between phonetic segments can be proposed and evaluated against psychophysical data. The objective is to simulate, e.g. perceptual distance data (Singh, 1971) and also category boundary shifts as acoustic cues are manipulated. For example, one can simultaneously manipulate several acoustic dimensions of synthetic speech-like stimuli (Massaro & Cohen, 1976; Stevens & Klatt, 1974), and try to account for cue tradeoffs. The simplest static metric might consist of summing the squares of the differences in dB across a set of critical-bandwidth filters. The simplest dynamic metric might consist of summing the static distances over time. If these and other kinds of simple metrics are falsified by perceptual studies, we would have evidence in favor of an intermediate level of analysis, perhaps one involving property detectors.

Several alternative talker-normalization procedures were proposed for LAFS. One involved continual modifications to the spectral templates used in the network on the basis of experience with the speech of the current talker. In order to retain this knowledge in long-term memory, a set of templates could also be saved for each familiar talker and for prototypical male and female talkers. Other techniques include computational procedures for modifying templates on the basis of vocal tract length estimates, average speech spectrum estimates, and average background noise spectrum estimates. All of these techniques are theoretically well motivated and thus potentially psychologically valid.

Experiments that examine the limits of listener's abilities to normalize for unusual spectral distortions may help to constrain the nature of the normalization process. For example, it seems that listeners are remarkably insensitive to changes in the relative amplitudes of spectral peaks caused by playing speech through fixed filters, but the limits of this ability have not been quantified when formant amplitude relations are disturbed dynamically.

Learning new words

In the event that the higher-level components determine that an unfamiliar word has been spoken [5], a phonetic representation of this speech interval [6] (produced by SCRIBER or perhaps by an augmented LAFS model) is recovered from a temporary phonetic buffer [7], submitted to morphological analysis [8], converted to a phonemic representation [9], and stored in the primary lexicon along with syntactic, semantic, and morphological properties. The new word is then incorporated into the spectral-sequence decoding network of LAFS through activation of procedures that include processing the abstract phonemic representation by a set of phonological rules [11], and expanding the resulting alternative phonetic forms into expected spectral sequences [12] that are then integrated with the LAFS network.

Addition of a new form to the network includes attaching it to the appropriate bound morphemes. For example, assuming that the talker has learned the pluralization rule of English in the form of a productive rule, she/he would have to look at the final phoneme of a new word, determine the distinctive feature categories to which it belonged, and define

a network path from the end of the new word to the appropriate plural morpheme sub-network[1]. The advantage of placing common bound morpheme suffixes, like the regular plural, in a special sub-network is to ensure that they be preceded only by the appropriate words.

Verification of top-down lexical predictions

There exist listening situations where specific words can be anticipated on the basis of situational context and prior dialog. In this case, one need not wait for LAFS to complete its bottom-up analysis, but, instead, the higher level components can scan the input as it arrives so as to make an early lexical decision and prepare for the next word.

The way that this is accomplished is not at all clear. LAFS might be modified to work interactively with syntactic and semantic routines (see Newell, 1978, for an extreme version of this alternative), or there may exist a special mechanism for top-down lexical prediction and verification. Both alternatives are shown in Fig. 5. A top-down lexical hypothesis [13] is sent to a special word verification component [13a] that scans the phonetic transcription produced by SCRIBER to compute a matching score that can be used to disconfirm the presence of an expected word even if the complete phonetic representation has not been received. This alternative has the advantage of preserving the autonomy of LAFS from syntactic/semantic influences (Forster, 1976). The second (non-exclusive) alternative, path [13b], is discussed below. Both are presented in dashed lines because of my limited interest in pursuing here the implications of this part of the model, and because there are other ways in which top-down lexical predictions could be verified.

Lexical ambiguity in noise As presently conceived, LAFS does not output a word until a certain amount of additional input is scanned. LAFS must wait long enough to resolve lexical ambiguities introduced when sub-words like "see" are contained in words like "cement", or alternative word sequences like "see mental loudmouths" and "cement allowed mountains" cover part of the phonetic input equally well. Unfortunately, the search space increases rapidly with delayed commitment. There is an exponential growth in alternative hypotheses with increased delay in commitment, so an autonomous LAFS probably cannot be allowed to defer decisions until several additional morphemes worth of input is scanned, but rather must try to make commitments with a minimum delay, no matter what the cost in performance.

Thus it is not at all clear how well an autonomous LAFS model (or any autonomous model of bottom-up lexical hypothesis formation) can perform in the context of a morpheme lexicon as large as in unconstrained English, or in the context of commonly encountered

[1]Derivation of phonological and morphological facts about new words appears to require a sophisticated network-building "demon". A computationally equivalent more plausible embodiment of precompiled knowledge might be to realize cross-word phonological recoding as a set of subroutines (rather than activate a demon to modify every word pair in the network that satisfies the preconditions for rule applicability). To learn a phonological rule would be equivalent to creating this subroutine. For example, when evaluating the score for a portion of the network corresponding to a postvocalic /st/ cluster for "list", one would scan the list of phonological rules that apply to postvocalic segments, detect the /st≠s/ → [s] rule, and thereby jump to the appropriate nodes of the network. The computational cost of scanning possible rules is offset by the need for a less-powerful demon. This is a standard tradeoff between computational speed and storage requirements that comes up often in computer programming, but we have no idea how the nervous system has solved the trade-off problem.

background noises. Speech can be understood in a moderately high noise background. The performance of an autonomous LAFS is likely to degrade significantly in noise (perhaps to a point where its output is essentially useless for moderate amounts of noise).

Given these observations, my present intuitions favor an interactive LAFS network in which particular lexical items or classes of items can be facilitated by a predictive syntactic/semantic module (path 13b of Fig. 5). The matching scores for these words are increased in proportion to the confidence of the top-down predictions, analogous to the Logogen model of Morton (1970). If message redundancy is high and these constraints can narrow the lexical search, fewer errors will be made in general, and performance should not degrade as rapidly in the presence of noise and ambiguity. The alternative ways to integrate acoustic and semantic cues to lexical items are discussed in greater detail in Marslen-Wilson & Welsh (1978) and in Morton & Long (1976). However, it is important to emphasize that the present paper is not concerned with this issue since the principles embodied in LAFS, i.e. lexical access from a network of spectral sequences for words, can be incorporated in either an autonomous module or an augmented interactive data structure.

In conclusion, steps have been taken in this section to transform LAFS into a plausible perceptual model. This has required few modifications to the LAFS network representation of knowledge or to the recognition strategy. However, additional components have been postulated for adding new words to LAFS, and it has been argued that a method is needed for applying predictive constraints within LAFS to narrow the search space.

Discussion

Why propose a new model of lexical access? Modeling efforts can serve several purposes: (1) to unify seemingly disparate facts into a cohesive theory, (2) to detect gaps in the knowledge available to support any model, and (3) to define testable alternatives to mechanisms described in previous models. The present model of the early stages of speech perception is far too speculative to qualify as a theoretical synthesis of available data. In the following paragraphs, the model is discussed with reference to the latter two objectives.

Precompiled knowledge

The most important idea to come out of recent efforts to build computerized speech understanding systems is that one can precompile detailed relations between acoustic and lexical events into an efficient high-performance decoding structure for speech analysis. Precompilation is a fundamental computational technique that is clearly potentially applicable in other domains. For example, during the development of speech production strategies, the motor commands to the articulators to realize a particular phonetic segment must be adjusting as a function of the current state of the articulatory apparatus, i.e. as a function of the previous phonetic segment. These adjustments can be learned as rules in a feature-based system, but they might also be precompiled and stored in the form of a network of motor instructions for all phonologically possible phonetic sequences of English.

Phonological rules

It is well known that words have different phonetic realizations depending on the sentence context. Phonological recoding seems to occur so as to simplify the task of the talker. Do these simplifications add a significant burden to the listener by increasing the ambiguity of speech? There is no doubt that ambiguity is generated since it is not possible

to write a unique inverse decoding rule for most phonological rules. How much ambiguity depends on whether the listener must consider all applicable inverse decoding rules at all phonetic positions in an unknown utterance, or whether a psychological equivalent of LAFS rule precompilation occurs.

It may be possible to determine experimentally whether phonological recodings are learned as generalized rules and precompiled into a decoding structure. For example, "list some" is likely to have a deleted [t], while "list one" may not. Reaction time experiments are needed to see if word strings subjected to phonological recoding and simplification require greater processing time. One would predict an insignificant increase in lexical access time if rule effects have been precompiled into a decoding structure like the LAFS spectral sequence decoding network. However, if rules are not precompiled and the effects of phonological recoding make lexical access considerably more difficult by introducing many alternative underlying phonetic strings, reaction time is likely to increase. Careful experimentation may reveal which, if any, phonological phenomena are detrimental to speech decoding.

Phonetic segments

The lexical recognition procedures of LAFS call into question the status of phonetic segments as units to be recognized during the bottom-up lexical hypothesis formation. The psychological reality of phonetic segments has been emphasized in the past on grounds of linguistic parsimony, in order to reduce the size of the knowledge store, in order to minimize the processing burden on the listener, and to serve as an interface between talking and listening. The model that we have proposed demands re-examination of these arguments.

A number of experiments have been devised to determine which units are involved in speech perception. For example, the LAFS model is consistent with reaction time data that indicate quicker reaction times for word processing over phoneme monitoring (Rubin, Turvey & Van Gelder, 1976; Savin & Bever, 1970). It has been suggested that the listener cannot access the phonemes that were used to recognize a word, but a clear alternative possibility is that word recognition does not usually involve phoneme recognition as an intermediate step. The absence of phonetic identification in LAFS is also consistent with the phonetic restoration effect (Warren, 1970) since all that is available from the output of LAFS is the best-scoring word. The LAFS model is also consistent with studies which indicate the perceptual migration of clicks to word boundaries (Ladefoged and Broadbent, 1960), with details of word advantage effects on a voice-onset-time continuum (Ganong, 1978), and with information theoretic arguments to the effect that listeners should not be required to make too many serial decisions per unit time (Miller, 1962).

Of course, we do not propose to discard the phonetic segment entirely in the perceptual model. Phonetic analysis skills are essential for adding words to the lexicon used for talking and for adding morphemes to LAFS. At an earlier developmental stage, phonetic analysis skills are probably essential for learning to talk (see below for a discussion of developmental issues).

Phonetic features

The model is intentionally provocative in its attempt to define a speech analysis system that does not make use of either simple acoustic property detectors or sophisticated phonetic feature detectors in any module (except that features serve as names for sets of

phonetic segments that participate in particular phonological rules of higher-level components). Phonetic feature analyzers are included in most current models of speech processing (Blumstein, *et al.*, 1977; Jakobson, Fant & Halle, 1953; Oden & Massaro, 1978; Pisoni & Sawusch, 1975; Pisoni, 1976). Evidence cited in support of feature analyzer concepts includes (1) the structure of perceptual confusion matrices (Miller & Nicely, 1955; Shankweiler & Studdert-Kennedy, 1967), (2) the phenomena of cross-adaptation (Cooper, 1978), (3) data on the perception of competing acoustic cues (Massaro & Cohen, 1976), and (4) categorical perception of consonants. Recent reviews of this literature (Ades, 1978; Ganong, 1979; Parker, 1977; Studdert-Kennedy, 1979) suggest that all of the effects noted are consistent with *acoustic* properties of phonetic segments taken as a whole, and that one is not *required* to conclude that segments are represented in terms of distinctive features at early stages of speech perception. There appear to be no data that would rule out either SCRIBER or LAFS as components of perceptual processing.

Along the same line, recent analyses of speech production error data suggest that segments are manipulated as unanalyzed wholes during the initial stages of speech production (Shattuck-Hufnagel & Klatt, 1979). The authors have shown that exchange errors of the type "*m*itt or *h*iss" involve the movement of whole segments rather than the movement of component distinctive features. Is it possible that distinctive features are nothing more than names for sets of phonetic segment types that participate in phonological rules during both speech production and perception? Distinctive feature theory will always serve as a set of unifying principles for the organization of languages and the definition of natural phonetic contrasts for humans to produce and perceive (Chomsky & Halle, 1968; Jakobson, Fant & Halle, 1953; Stevens, 1972*a*), but it really has not been established that this representation is employed by the language user at acoustic and articulatory levels.

Perhaps the promulgation of LAFS as a creditable psychological model will stimulate the design of new types of experiments that can distinguish between feature-based, segment-based, and word-based accounts of acoustic-phonetic processing and lexical access. While I offer below a few suggestions on this point, note the cautionary words of Licklider (1952) who pointed out that certain classes of feature-analysis systems and template-based systems are functionally equivalent in the restricted mathematical sense that either can compute the same input-output transformation.

The model described here represents speech by sequences of acoustic events. Identification is determined by how well the input matches individual category prototypes. Our approach is thus similar to one advocated by Oden and Massaro (1978), except that our prototypes are defined in terms of sequences of spectral templates, rather than in terms of the outputs of feature detectors. In order to distinguish between a spectral-sequence model and a feature-detector model, two hypotheses must be tested: (1) whether metrics can be developed to predict psychophysical similarity between phones on the basis of general spectral and temporal properties, and (2) whether rate of spectral change is an important *independent* variable in speech perception.

Stevens (1972*a*) and others have argued that certain properties (such as the distinction between a rapid onset versus a gradual onset, or simultaneous versus sequential acoustic onsets) are natural psychophysical dimensions along which languages divide their phonetic inventories. It will be interesting to see whether the spectral representation and distance metrics proposed here can account for these and other natural phonetic contrasts, or whether special property detectors (or prototypes more complex than a sequence of spectral templates) must be postulated.

Relation to other models of speech perception

Analysis by synthesis An analysis-by-synthesis strategy was first formulated at the level of phonetic segments and features in order to overcome the non-invariance problem (Halle & Stevens, 1962; Stevens & Halle, 1964; Stevens, 1972*b*). The theoretical advantages of analysis-by-synthesis concepts applied at the lexical level have been hinted at in the section entitled "The problem". Since the literature does not contain a description of how such a lexically-based model might work, the following paragraphs describe one possible realization.

A block diagram of an Analysis-by-synthesis model is shown in Fig. 6. The speech waveform corresponding to an unknown utterance arrives at the left in the diagram and the best-scoring word string leaves at the right. It is assumed that peripheral cochlear spectral analysis and central neural processing result in a spectral representation for the incoming speech. This representation is placed in a temporary "echoic" memory store having characteristics that are not unlike those listed earlier in Table III.

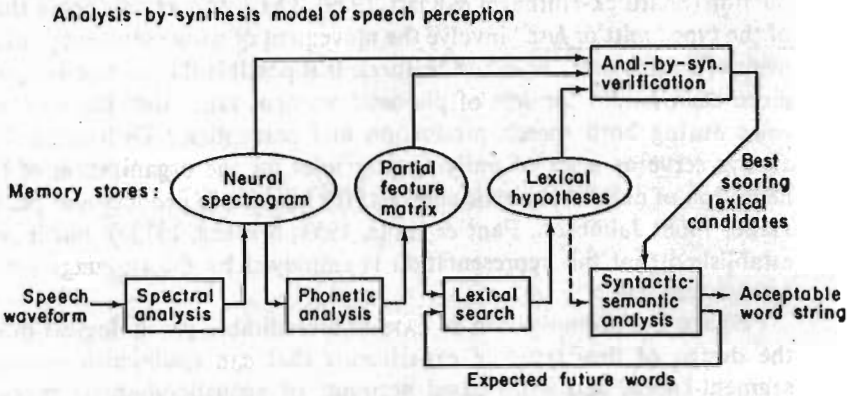

Figure 6. Simplified block diagram of an analysis-by-synthesis model of speech perception. Memory buffers are represented by ovals, processes by rectangles, and information flow by arrows.

Results of backward recognition masking experiments suggest that only about 200–300 ms of speech can be retained in this memory store at any given time (Massaro, 1975). Certain kinds of transient acoustic information contained in the neural spectrogram may be lost unless they are quickly recoded into phonetic form, although prosodic information such as aspects of the fundamental frequency contour, intensity envelope, segmental durations, and vowel quality are presumably recoded and placed in a short-term memory store to permit relative comparisons across greater time spans of the input.

Phonetic feature analyzers read off information from the neural spectrogram. The phonetic feature detectors are thought to place their outputs in a matrix, where the columns represent a division of the time dimension into phonetic segments and the rows represent different phonetic features. Each entry in the matrix indicates whether or not a particular segment has a given feature, or is left unspecified. The phonetic feature specification for a segment may be incomplete for two reasons. Either sloppy articulation or environmental noise has led to an ambiguity, or some decisions involving complex context-dependent acoustic-phonetic relations are not attempted in the preliminary phonetic transcription. If specified, features are usually thought to be binary, representing the presence or absence of some attribute in the partial feature matrix representation of

Fig. 6. Alternatively, there may be an advantage in using continuous scales to represent the degree of confidence in each feature decision in order to provide more information for lexical hypothesis formation.

Information drawn from the phonetic feature matrix is used to search through the lexicon for possible sentence-initial words (or morphemes). Lexical candidates expected on the basis of the situation and previous dialog may be "facilitated" so that even if the feature match is not perfect, these highly probable words will be included in the lexical hypothesis memory store.

Due to the errorful and incomplete nature of the preliminary phonetic feature representation of the input speech, a rather large number of lexical candidates are likely to be found—perhaps an unreasonably large number (Klatt & Stevens, 1973). Syntactic/semantic routines would be hard-pressed to choose the correct initial word and succeeding words given this amount of ambiguity. Thus the model of Fig. 6 includes an analysis-by-synthesis component that accepts both bottom-up and top-down lexical hypotheses and returns to the acoustic and phonetic data to verify the presence of details to be expected given the whole word (Klatt, 1975). Verification takes into account not only the acoustic manifestations of phonetic expectations, but also perturbations caused by the prosodic cues that are expected given the lexical stress pattern and syntactic category(s).

Once the best-scoring sentence-initial lexical candidate has been accepted by the syntactic/semantic module, the whole process is iterated, starting from the position in the partial feature matrix where the first word ended. [The limited capacity of the partial feature matrix memory store probably means that sentence processing must proceed in near real time.] The words of a candidate sentence must also fit together properly. Given a knowledge of the previous word in an hypothesized sentence fragment, phonological rules contained in the verification component can be used to specify permitted segmental recoding across word boundaries and thus determine whether the word pair is compatible with the acoustic data. Significant cross-word phonological and coarticulatory interactions are common in sentences. Words rarely appear in a canonical form specified by the lexicon, and some means of dealing with word variability is essential.

As the sentence is elaborated, greater reliance is made on predictions given by syntactic/semantic expectations. These expectations might be used to order the lexical candidates in the lexical hypothesis memory store, or to indicate to the verification component that verification need not be as detailed, or to indicate that not all candidates need evaluation if the expected word scores reasonably well. While the process should rarely reach a dead end (signifying an incorrect analysis), when it does, memory stores are searched to try to backtrack to the next-best alternative partial sentence.

The block diagram of Fig. 6 is just a framework or philosophy for speech understanding, not a complete model. None of the components have been fleshed in with very great detail, particularly the partial feature analysis stage and how the lexicon is searched to find lexical candidates bottom up. It is hoped that this paper will stimulate efforts to refine feature-based analysis-by-synthesis theories of speech perception that deal with the eight problem areas of Table II, and that the paper will help to generate critical experiments to determine whether feature-based analysis-by-synthesis theories are more realistic models of human sentence perception than the model of Fig. 5.

Analysis by synthesis is a powerful (though expensive) method of weeding out false word candidates. The power derives from the fact that segment durations, intensity contour and fundamental frequency contour must make sense given the large number of factors that contribute to the expected patterns. Similarly, all of the acoustic cues that

contribute to phonetic distinctions make more sense when a lexical hypothesis is under consideration. However, these advantages can be incorporated directly in a LAFS structure. Precompilation of knowledge is an attractive form of analysis by synthesis; the LAFS strategy permits the evaluation of far more lexical possibilities in parallel than if a single analysis-by-synthesis module were to be activated by one top-down lexical hypothesis at a time.

For familiar words, precompilation of acoustic-phonetic and phonological knowledge is a big winner. However, there probably are situations where precompilation is not the best solution, and other strategies (such as analysis-by-synthesis) may be invoked by the listener. For example, when listening to speakers with foreign accents or unfamiliar dialects, understanding improves when one has deduced a theory of the phonetic recoding, but it may not be desirable to compile this special knowledge into LAFS.

The motor theory Liberman *et al.* (1967) and Studdert-Kennedy, Liberman, Harris & Cooper (1970) have argued that the acoustic encoding of phonetic elements in spoken utterances is so complex that it requires a special decoder. The decoder attempts to determine an articulatory sequence that could have produced the observed acoustic pattern. This motor theory of speech perception postulates an intermediate articulatory representation between the acoustic data and the phonetic interpretation. The need for referral to a motor component in speech perception has been further elaborated by Liberman and Studdert-Kennedy (1978). A model based on this philosophy has never been specified in very great detail. However, no matter what form the model takes, in principle all of the complex acoustic-articulatory-phonetic relations implicit in a motor theory can be precompiled into a network of expected spectra for each phonetic transition. If this were done, the resulting network would hopefully be indistinguishable from SCRIBER. In a sense, SCRIBER can be viewed as a computationally equivalent passive form of an active motor theory.

Is it possible to reformulate SCRIBER or LAFS to include an intermediate articulatory level? An acoustic-to-articulatory transformation can be computed, at least approximately (Atal, 1975), and could form an intermediate step in a modified LAFS model of lexical access in which states become articulatory configurations. Such a system would have no computational advantage over the present LAFS, since both can compute the same decoding transformation with about the same computational cost, but an articulatory LAFS might be more suitable for the interface between speech perception and speech production. A lexical-access-from-articulatory-sequences model seems worthy of investigation for this reason, but it does not appear that a lexical decoding network based on articulation could be used simultaneously for speech production. The representation is not sufficiently abstract—all of the coarticulatory and phonological details that are a consequence of low-level articulatory dynamics are already represented in the decoding network, whereas the lexicon used for talking is almost certainly phonemic (Fromkin, 1971; Shattuck-Hufnagel & Klatt, 1979).

The logogen Morton (1970) has proposed a model of lexical access in which each word in the lexicon has an associated "logogen" that specifies definining characteristics of the word along various acoustic and semantic dimensions. If enough of these features are satisfied during sentence processing, the word is recognized. A Logogen model can account for many kinds of experimentally determined interactions between acoustic and semantic cues to lexical identity (Morton & Long, 1976) by postulating a threshold mechanism whereby top-down expectations can push a word over threshold before all acoustic cues are seen. There is no description of how acoustic cues are processed in a

Logogen model. The model is clearly compatible with the general acoustic-to-lexical analysis framework outlined here. In fact, LAFS could be considered as a more specific characterization of how the bottom-up part of a Logogen model would function.

A Logogen model has some difficulty in accounting for details of word reaction time data (Marslen-Wilson & Welsh, 1978). However, the autonomous LAFS model is similarly deficient in that there must be a delay before identifying each word (because one has to wait until at least the next word is over to be confident of selecting the best path in the beam search) and this is also inconsistent with reaction time data.

Context-sensitive allophones Wickelgren (1969, 1976) proposed a theory of speech perception in which a large set of context-dependent allophones are used to derive a phonetic representation for an unknown utterance. For each phoneme X, a set of context-dependent allophones aXb were defined for all possible preceding phonemes, a, and all possible following phonemes, b. While such an approach solves the non-invariance problem, it does not address most of the other problems listed in Table I. The solution to the non-invariance problem proposed by Wickelgren is slightly more powerful than the diphone approach (and considerably more costly in terms of number of basic elements to be recognized). It suggests a way in which SCRIBER and LAFS can be improved in those cases where diphones do not capture all of the context dependency of speech. If, for example, the acoustic characteristics of /I/ in a word like "will" cannot be predicted from diphones obtained from "*wi*th" and "h*ill*" because the /w/ and /l/ collectively velarize the /I/ to a greater extent, then a special context-dependent allophone, or "triphone", can be defined in terms of a sequence of spectral templates and placed in the network in place of the two concatenated diphones.

Sequential word recognition Cole & Jakimik (1978) have used a mispronunciation detection task to show that sentence perception generally involves the direct left-to-right decoding of words, one after the other. The advantage of such a strategy is that the end of one word defines the beginning of the next word in time, thus reducing the potential ambiguity of looking for words starting at other phonetic positions in the sentence. LAFS incorporates this advantage of direct left-to-right processing of a sentence, and it adds the further advantage of being able to deal with phonological recoding across word boundaries.

The hearsay II blackboard Hearsay II is one of several computer-based speech understanding systems developed during the ARPA speech understanding project (for a review, see Klatt, 1977). The Hearsay II blackboard model of speech perception is described by Erman (1978). In this model, or framework for speech understanding, a set of knowledge sources work asynchronously toward the decoding of a sentence by taking their input from a common blackboard and placing the results of their analyses back on the blackboard. LAFS could function as a component of such a blackboard model. Alternatively, the theories being considered in parallel by LAFS could be placed on the blackboard for examination by other modules, even before LAFS has made a final decision. The latter possibility forms the basis for a number of attractive alternative models of speech perception, but they all will have to face such inherent problems as how to schedule activity among the modules that interact with the blackboard, and how to deal with the halting problem (Reddy, 1978).

Is speech special?

If our model is correct, one need not postulate the existence of innate feature detectors sensitive only to the phonetic contrasts of spoken language (Eimas & Miller, 1978).

Instead, certain natural discriminations (so natural as to be made by infants) would be the consequence of properties of the spectral sequence representation of auditory signals.

However, speech could be special in several other respects. For example, speech stimuli may be distinguished from non-speech stimuli because they are the only signals that receive high-enough matching scores in the outputs of LAFS and SCRIBER to be processed as language. Also, the steps involved in constructing and augmenting a LAFS decoding network are complex. Is LAFS representative of general cognitive strategies (in which precompiled knowledge networks play a prominent role) or does speech acquisition require the postulation of special innate structures for the development of LAFS and supporting higher-level components?

Developmental issues

The earliest representation of words by an infant is probably in the form of a crude direct encoding of what appears on the (hard-wired?) neural spectrogram. Perhaps only a few of the most prominent spectral details within a word are remembered at first. The actual memory representation for words may thus be quite similar to a LAFS sequence of spectra representation right from the beginning. On the basis of further experience, spectral details are filled in, but only when needed to differentiate between new words.

In order to learn to talk, a phonemic analysis of the input speech must then be discovered by relating the processes of listening (acoustic events) and talking (articulatory commands). The creation of a phonemic talking lexicon is no doubt facilitated by the presence of partial acoustic-phonetic invariance. It seems that many invariant (or nearly invariant) cues must be present if the child is to discover the phonemic structure of his/her native language. However, according to the views expressed here, the acquisition of phonemic analysis capabilities and of a phonemic talking lexicon does not lead to any fundamental changes in bottom-up lexical access of familiar words via LAFS.

Relations between the two representations used for talking and listening are then internalized by associating spectral sequences with each phoneme or phoneme pair so as to create the diphone dictionary required for top-down augmentations of LAFS. The final steps needed to acquire an adult-like LAFS decoding network are the acquisition of morphological decomposition skills and the discovery of how word sequences are modified by phonological rules. The perceptual model thus acknowledges the psychological reality of linguistic units and rules that never appear explicitly in LAFS. Just how sophisticated these processes are, however, is a subject for experimentation (Ohala, 1974).

Elaboration of a concrete model of speech acquisition along these lines would be an important contribution to the general theory of speech perception. Hopefully, many testable alternatives can be isolated by comparing this account of language development with other current theories.

The validation issue The presence of both LAFS and SCRIBER in the proposed perceptual model makes it much more difficult to determine the psychological validity of either. Depending on the perceptual task (nonsense-syllable identification, repeated listening to the same pair of words, or listening to unpredictable sentences made up of familiar words), the listener may engage either or both mechanisms. Similarly, analysis by synthesis or another form of top-down verification employing generative rules that are computed on the fly may be invoked when listening to some speakers. Nevertheless, I believe that the efficient decoding of normal conversational speech depends critically on mechanisms found in LAFS, whether or not these are the only mechanisms used in speech perception.

Conclusions

The perceptual model shown in the bottom half of Fig. 5 has been proposed and compared with a number of alternative models. I have established the theoretical advantages of a spectrally based decoding network approach to speech analysis, and have suggested several kinds of experiments that might settle the issues that have been raised concerning its perceptual reality. The essential features of the model are (a) precompilation of phonological rules that describe phonetic recoding of words in sentences so as to avoid having to consider application of inverse rules indiscriminately, (b) no calculation of a phonetic level of representation during lexical search because calculation of such an intermediate representation must introduce errors (due in part to the greater number of alternatives in a phonetic transcription and in part to an inability to interpret durational and FO cues to segmental contrasts) thus violating the principle of delayed commitment, and (c) representation of acoustic-phonetic knowledge in terms of sequences of spectra for each possible phonetic transition rather than postulating the existence of invariant attributes for phones or the existence of low-level property detectors and phonetic feature detectors until such time as simpler assumptions are proven unworkable. This model, summarized in Fig. 5, is offered as the most complete, most simply structured current theory of the initial stages of acoustic-phonetic analysis and lexical search.

Preparation of this manuscript was supported by an NIH grant. My sincere thanks go to R. Cole, A. Liberman, M. Liberman, D. Pisoni, R. Reddy, B. Repp, and K. Stevens for numerous suggestions for improvements to an earlier draft. I alone take responsibility for the views expressed here.

References

Ades, A. E. (1978). Theoretical notes: vowels, consonants, speech and nonspeech. *Psychological Review* 84, 524–30.

Allen, J. (1973). Speech synthesis from unrestricted text. In *Speech Synthesis* (J. L. Flanagan & L. R. Rabiner, eds). Stroudsberg, PA: Dowden, Hutchinson & Ross.

Atal, B. S. (1975). Towards determining articulator positions from the speech signal. In *Speech Communication* (G. Fant, ed.), Vol. 1, pp. 1–9. Stockholm: Almqvist & Wiksell.

Blumstein, S. E., Stevens, K. N. & Nigro, G. N. (1977). Property detectors for bursts and transitions in speech perception. *Journal of the Acoustical Society of America* 61, 1301–13.

Carlson, R., Fant, G. & Granstrom, B. (1975). Two-Formant Models, Pitch, and Vowel Perception. In *Auditory Analysis and Perception of Speech* (G. Fant & M. A. A. Tatham, eds). New York: Academic Press.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Cohen, P. S. & Mercer, R. L. (1975). The phonological component of an automatic speech recognition system. In *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium* (D. R. Roddy, ed.), 275–320. New York: Academic Press.

Cole, R. & Jakimik, J. (1978). A model of speech perception. In *Perception and Production of Fluent Speech* (R. Cole, ed.). Hillsdale, NJ: Erlbaum Assoc.

Cole, R. A. & Scott, B. (1974). Toward a theory of speech perception. *Psychological Review* 81, 348–74.

Cole, R., Rudnicky, A., Zue, V. and Reddy, D. R. (1978). Speech as patterns on paper. In *Perception and Production of Fluent Speech* (R. Cole, ed.). Hillside, NJ: Erlbaum Assoc.

Cooper, W. E. (1978). *Speech Perception and Production: Selected Studies on Adaptation*. Cambridge, England: Cambridge University Press.

Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Accustical Society of America* 27, 769–73.

Egan, J. P. & Hake, H. W. (1950). On the masking pattern of a simple auditory stimulus. *Journal of the Acoustical Society of America* 22, 622–30.

Eimas, P. D. & Miller, J. L. (1978). Effects of selective adaptation on the perception of speech and visual patterns: evidence for feature detectors. In *Perception and Experience* (R. D. Walk & H. L. Pick, eds). pp. 307–45. New York: Plenum Press.

Erman, L. (1978). The HEARSAY-II speech understanding system. In *Trends in Speech Recognition* (W. A. Lea, ed.). New York: Prentice-Hall.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G. (1974). *Speech Sounds and Features*. Cambridge, MA: MIT Press.
Flanagan, J. L. (1957). Estimates of the maximum precision necessary in quantizing certain dimensions of vowel sounds. *Journal of the Acoustical Society of America* **29**, 533–4.
Forster, K. I. (1976). Accessing the mental lexicon. In *New Approaches to Language Mechanisms*. (R. J. Wales & E. C. T. Walker, eds). Amsterdam: North-Holland.
French, N. R. & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America* **19**, 90–119.
Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language* **47**, 27–52.
Fujimura, O. & Lovins, J. B. (1978). Syllables as concatenative phonetic units. In *Syllables and Segments*. (A. Bell & J. B. Hooper, eds).
Ganong, F. (1978). A word advantage in phoneme boundary experiments. *Journal of the Acoustical Society of America* **63**, Suppl. 1, S20 (A).
Ganong, F. (1979). Dichotic feature recombination errors and distinctive features. Unpubl. manu.
Gay, T. (1977). Articulatory movement in VCV sequences. *Journal of the Acoustical Society of America* **62**, 183–193.
Haggard, M., Ambler, X. & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America* **47**, 613–17.
Halle, M. & Stevens, K. N. (1962). Speech recognition: a model and a program for research. *IRE Transactions on Information Theory IT-8*, 155–9.
Houtgast, T. (1974). Auditory analysis of vowel-like sounds. *Acoustics* **31**, 320–4.
Jakobson, R., Fant, G. & Halle, M. (1953). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press.
Kiang, N., Watanabe, T., Thomas, E. & Clark, L. (1965). *Discharge Patterns of Single Fibres in the Cat's Auditory Nerve*. Cambridge, MA: MIT Press.
Klatt, D. H. (1975). Word verification in a speech understanding system. In *Speech Recognition: Invited Papers Presented at the* 1974 *IEEE Symposium* (D. R. Reddy, ed.), pp. 321–41. New York: Academic Press.
Klatt, D. H. (1976*a*). A digital filter bank for spectral matching. In *Conference Record of the* 1976 *IEEE International Conference on Acoustics Speech and Signal Processing* (C. Teacher, ed.). Philadelphia, PA. (IEEE Catalog No. 76CH1067-8 ASSP), 537–40.
Klatt, D. H. (1976*b*). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America* **59**, 1208–21.
Klatt, D. H. (1977). Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America* **62**, 1345–66.
Klatt, D. H. (1978*a*). SCRIBER and LAFS: two new approaches to speech analysis. In *Trends in Speech Recognition* (W. A. Lea, ed.). New York: Prentice-Hall.
Klatt, D. H. (1978*b*), Analysis and synthesis of consonant-vowel syllables in English. *Journal of the Acoustical Society of America* **64**, Suppl. 1, S43 (A).
Klatt, D. H. (1979), Synthesis by rule of segmental durations in English sentences. In (B. Lindblom & S. Ohman, eds). *Frontiers of Speech Communication Research* eds. New York: Academic Press.
Klatt, D. H. & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In *Structure and Process in Speech Perception* (A. Cohen & S. G. Nooteboom, eds), pp. 69–89. New York: Springer-Verlag.
Klatt, D. H. & Stevens, K. N. (1973). On the automatic recognition of continuous speech: implications of a spectrogram-reading experiment. *IEEE Transactions on Audio and Electroacoustics AU-21*, 210–17.
Klovstad, J. W. (1978). Computer-automated speech perception system. Ph.D. Dissertation. MIT, unpubl.
Ladefoged, P. & Broadbent, D. E. (1960). Perception of sequence in auditory events. *Quarterly Journal of Experimental Psychology* **13**, 162–70.
Lea, W. A. (1973). Segmental and suprasegmental influences on fundamental frequency contours. In *Consonant Types and Tone* (L. Hyman, ed.). Southern California Occasional Papers in Linguistics, No. 1.
Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
Lehiste, I. & Shockey, L. (1972). On the perception of coarticulation effects in english VCV syllables. *Journal Speech and Hearing Research* **15**, 500–6.
Leshowitz, B. & Cudahy, E. (1975). Masking patterns for continuous and gated sinusoids. *Journal of the Acoustical Society of America* **58**, 235–42.
Liberman, A. M., Cooper, F. S., Shankweiler, D. S. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* **74**, 431–61.
Liberman, A. M., Delattre, P., Gerstman, L. & Cooper, F. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology* **52**, 127–37.
Liberman, A. M. & Studdert-Kennedy, M. (1978). Phonetic perception. In *Handbook of Sensory Physiology*, Vol. VIII (R. Held, H. Leibowitz & H.-L. Teuber, eds). Heidelberg: Springer-Verlag.

Licklider, J. C. R. (1952). On the process of speech perception. *Journal of the Acoustical Society of America* **24**, 590–4.
Lieberman, P. (1967). *Intonation, Perception, and Language*. Cambridge, MA: MIT Press.
Lindblom, B. (1978). Phonetic aspects of linguistic explanation. *Studia Linguistica* (in press).
Lowerre, B. T. (1976). The HARPY speech recognition system. Ph.D. Dissertation, Carnegie-Mellon Univ., unpublished.
Lowerre, B. T. (1977). Dynamic speaker adaptation in the HARPY speech recognition system. In *Conference Record of the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing*, (H. F. Silverman, ed.), Hartford, IEEE Catalog No. 77CH1197-3 ASSP.
Lowerre, B. & Reddy, D. R. (1978). The HARPY speech understanding system. In *Trends in Speech Recognition* (W. A. Lea, ed.). New York: Prentice-Hall.
Marslen-Wilson, W. D. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* **100**, 29–63.
Massaro, D. M. (1975). Backward recognition masking. *Journal of the Acoustical Society of America* **58**, 1059–65.
Massaro, D. M. & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset Time to the /zi/–/si/ distinction. *Journal of the Acoustical Society of America* **60**, 704–7.
Medress, M. (1969). Computer recognition of single-syllable English words. Ph.D. Dissertation. MIT, unpublished.
Miller, G. A. (1962). Decision units in the perception of speech. *IRE Transactions on Information Theory IT-8*, 81–3.
Miller, G. A. & Nicely, P. E. (1955). Analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* **27**, 338–53.
Morton, J. (1970). A functional model for memory. In *Models of Human Memory*, (D. A. Norman, ed.). New York: Academic Press.
Morton, J. & Long, J. (1976). Effect of word transition probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior* **15**, 43–51.
Newell, A. (1978). HARPY, production systems, and human cognition. In *Perception and Production of Fluent Speech* (R. Cole, ed.). Hillsdale, NJ: Erlbaum Assoc.
Oden, G. C. & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review* **85**, 172–91.
Ohala, J. J. (1974). Experimental historical phonology. In *Historical Linguistics II: Theory and Description in Phonology* (J. M. Anderson and C. Jones, eds). Amsterdam: North-Holland, 353–89.
Oshika, B., Zue, V. W., Weeks, R. V., Neu, H. & Aurbach, J. (1975). The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-23*, 104–12.
Parker, F. (1977). Distinctive features and acoustic cues. *Journal of the Acoustical Society of America* **62**, 1051–4.
Peterson, G. E., Wang, W. & Sivertsen, E. (1958). Segmentation techniques for speech synthesis. *Journal of the Acoustical Society of America* **30**, 739–42.
Pisoni, D. B. (1976). Speech perception. In *Handbook of Learning and Cognitive Processes* (W. K. Estes, ed.). Hillsdale, NJ: Erlbaum Associates.
Pisoni, D. B. & Sawusch, J. R. (1975). Some stages of processing in speech perception. In *Structure and Process in Speech Perception* (A. Cohen & S. G. Nooteboom, eds). New York: Springer-Verlag, 16–35.
Plomp, R. & Mimpen, A. M. (1968). The ear as a frequency analyzer II. *Journal of the Acoustical Society of America* **43**, 764–8.
Port, R. F. (1979). Influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, **7**, 45–56.
Reddy, D. R. (1978). Machine models of speech. In *Perception and Production of Fluent Speech* (R. Cole, ed.). Hillsdale, NJ: Erlbaum Assoc.
Rubin, P., Turvey, M. & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken non-words. *Perception and Psychophysics* **19**, 394–8.
Savin, H. B. & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior* **9**, 295–302.
Searle, C., Jacobson, J. Z. & Rayment, S. G. (1979). Stop consonant discrimination based on human audition. *Journal of the Acoustical Society of America* **65**, 799–809.
Shankweiler, D. & Studdert-Kennedy, M. (1967). Identification of consonants and vowels presented to left and right ears. *Journal of Experimental Psychology* **19**, 59–63.
Sharf, B. (1970). Critical bands. In *Foundations of Modern Auditory Theory, Vol.* 1 (J. V. Tobias, ed.), pp. 157–202. New York: Academic Press.
Shattuck-Hufnagel, S. R. & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior* **18**, 41–56

Singh, S. (1971). Perceptual similarities and minimal phonemic differences. *Journal of Speech and Hearing Research* **14**, 113–24.

Stevens, K. N. (1972*a*). The quantal nature of speech: evidence from articulatory-acoustic data. In *Human Communication: A Unified View* (E. E. David & P. B. Denes, eds). New York: McGraw-Hill.

Stevens, K. N. (1972*b*). Segments, features, and analysis by synthesis. In *Language by Eye and by Ear* (J. F. Kavanaugh & I. G. Mattingly, eds). pp. 47–55. Cambridge, MA: MIT Press.

Stevens, K. N. (1972*c*). Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. *Proc. Seventh Int. Congress of Phonetic Sciences* (A. Rigault & R. Charbonneau, eds), pp. 206–232. The Hague: Mouton.

Stevens, K. N. (1975). On the potential role of property detectors in the perception of consonants. In *Auditory Analysis and the Perception of Speech* (G. Fant & M. A. A. Tatham, eds). New York: Academic Press.

Stevens, K. N. & Halle, M. (1964). Remarks on analysis by synthesis and distinctive features. In *Proc. of the AFCRL Symposium on Models for the Perception of Speech and Visual Form* (W. Wathen-Dunn, ed.). Cambridge, MA: MIT Press.

Stevens, K. N. & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America* **55**, 653–8.

Studdert-Kennedy, M. (1976). Speech perception. In *Contemporary Issues in Experimental Phonetics* (N. J. Lass, ed.). New York: Academic Press.

Studdert-Kennedy, M. (1979). *Language and Speech* (in press).

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S. & Cooper, F. S. (1970). Motor theory of speech perception: a reply to Lane's critical review. *Psychological Review* **77**, 234–49.

Wakita, H. & Kasuya, H. (1977). A study of vowel normalization and identification in connected speech. In *Conference Record of the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing* (H. F. Silverman, ed.), pp. 417–27. Hartford: IEEE Catalog No. 77CH1197-3 ASSP.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science* **167**, 392–3.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review* **76**, 1–15.

Wickelgren, W. A. (1976). Phonetic coding and serial order. In *Handbook of Perception*, Vol. VII. New York: Academic Press, 227–64.

Weeks, R. V. (1974). Predictive syllable mapping in a continuous speech understanding system. In *Contributed Papers of the IEEE Symposium on Speech Recognition* (L. D. Erman, ed.). Carnegie-Mellon University: IEEE Catalog No. 74CH0878-9 AE.

Wolf, J. J. & Woods, W. A. (1978). The HWIM speech understanding system. In *Trends in Speech Recognition* (W. A. Lea, ed.). New York: Prentice-Hall.

Woods, W. A. & V. Zue (1976). Dictionary expansion via phonological rules for a speech understanding system. In *Conference Record of the 1976 IEEE International Conference on Acoustics Speech and Signal Processing* (C. Teacher, ed.), pp. 561–4. Philadelphia, PA: IEEE Catalog No. 76CH1067-8 ASSP.

Zue, V. W. (1976). Acoustic characteristics of stop consonants: a controlled study. *Lincoln Laboratory Technical Report No.* 523, Cambridge, MA: MIT.

Zue, V. W. & Schwartz, R. (1978). Acoustic processing and phonetic analysis. In *Trends in Speech Recognition* (W. A. Lea, ed.). New York: Prentice-Hall.

Zwicker, E., Terhardt, E. & Paulus, E. (1979). Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America* **65**, 487–498.