CHAPTER 3

AUDITORY PREPROCESSING AND RECOGNITION OF SPEECH

Roy D. Patterson and Anne Cutler

MRC Applied Psychology Unit,
15 Chaucer Road,
Cambridge CB2 2EF,
England

INTRODUCTION

For some time now, research on automatic speech recognition (ASR) has been largely concerned with what might be called the signal processing approach, in which the recognition of speech by machines was viewed as an information processing problem, quite distinct from the problem of how humans recognise speech. The signal-processing approach has had considerable success in the sense that it has produced a succession of special purpose devices that can recognise speech provided the vocabulary and the number of speakers is limited. It has not, however, led to the development of a general purpose ASR machine that can handle continuous speech from an arbitrary group of speakers using the vocabulary typical of normal conversation. It is also the case that the performance of current systems falls away rapidly when they are required to operate in noisy or reverberant environments.

The question then arises as to how best to proceed in the pursuit of the general purpose ASR machine. Some speech scientists argue that we should continue with the signal processing approach; that the quickest and surest route to the general machine is to refine the existing techniques and algorithms. An excellent review of the signal processing approach is presented in Bristow (1986). Others argue that current systems have inherent limitations that cannot be overcome within the signal processing framework, and that we must begin again with new concepts and processes. A portion of the latter group feel that the best way to proceed is to determine how humans process speech sounds and to develop a functional model of the human hearing and speech systems, the only speech recogniser with proven ability. An ASR machine which divides the problem up into the same sub-processes as the human brain, which provides some equivalent of the processing observed at each stage, and which performs the transformations in the same order, seems more likely to be successful than one which pays less attention to the human solution. This is the cognitive psychology approach, in contrast to the signal processing approach, and it is this approach that is the topic of the current chapter.

The psychological approach has the advantage of face validity; in the longer term it is bound to succeed. The problem is that we do not currently understand human speech processing well enough to assemble a complete functional model of the system, and even if we did, it would be too large to serve as the basis for a commercial ASR machine at this point in time. The purpose of this chapter, however, is not to explain how the psychological approach can solve all of the problems of speech recognition either now or in the near future. Rather its purpose is to point to the limitations of the signal-processing approach that have led to the re-emergence of the psychological direction in speech recognition, and to highlight some of the advances achieved by and projected for the psychological approach.

It is important to note that we are making the distinction between the signal-processing approach and the cognitive-psychology approach primarily in order to delimit the topic of this chapter. Like most dichotomies, it is not a hard and fast distinction. Furthermore, it is undoubtedly the case that both approaches will play a role in the development of ASR machines, along with others that do not even appear in the chapter. In making the distinction we simply intend to focus attention on a new direction in speech research and to indicate the origins and predilections of the scientists involved. The chapter is also restricted in terms of the portion of the speech problem with which it is concerned. It covers the processing of speech from the initial reception of an acoustic signal by the peripheral auditory system to the location in memory of a corresponding stored representation. It does not cover any higher-level processing such as the selection between alternative meanings for a homophone, contextual facilitation effects, syntactic evaluation, or integration into semantic context.

The research we will summarise falls into three parts: auditory perception which has traditionally been the province of psychoacousticians, word recognition which has traditionally been studied by psycholinguists, and the interface between the two which is essentially a new area of research. Parts 1 and 2 of this chapter outline the current research issues in auditory perception and word recognition, respectively. The description of the interface is deferred until Part 3, despite its logical position between hearing and speech, because it is qualitatively different Whereas auditory perception and word recognition are established research areas that can be reviewed in a straightforward way, interfacing models of hearing and speech is a new speculative venture which is currently characterised by small scale demonstrations of promising leads rather than proven large scale systems. A brief description of the interface problem is presented in the next subsection of the Introduction. The final subsection presents an extended example of one of the problems with the signal-processing approach to illustrate the motivation for returning to auditory models as preprocessors for auditory speech recognition.

A. Interfacing Auditory Models with Speech Models

Although speech sounds are a subset of auditory perceptions, there has been surprisingly little interaction between psychoacousticians and psycholinguists over the years. One of the main problems is that the two groups work with very different representations of sounds; the psychoacousticians represent speech, like other sounds, as arrays of filtered waveforms, whereas psycholinguists have tended to use phonetic codes, or some other discrete representation of sounds. The auditory models are massively parallel with from 30 to 300 channels, the parallelism continues through a number of auditory processing stages, and the reduction to a stream of auditory sensations occurs late in the system if it occurs at all. Speech models, in contrast, typically begin with relatively simple spectral analyses and reduce the parallel output of the spectral analysis to a serial string of speech features as early in the system as possible. Thus, the two types of model have different internal representations, involving vastly different data rates, throughout the majority of the processing stages, and it has not been possible to assemble an integrated model in which the output of an auditory front-end constructed by a psychoacoustician is used as the input to a speech processor constructed by a psycholinguist.

The current chapter provides an unfortunate example of the problem of differing internal representations. The first and second parts were written by a psychoacoustician and a psycholinguist, respectively, and despite our efforts to integrate them, the continuity of the chapter is severely disrupted by the differences in the representations used in the two parts. The contrast is useful, however, insofar as it shows the enormity of the speech recognition problem when one attempts to assemble a complete cognitive psychological representation of the process.

Recently, the situation has begun to change, and in an interesting way. Psychoacousticians and speech scientists with a psychological orientation have begun developing spectre-temporal auditory models to simulate the neural firing patterns produced in the auditory system by complex sounds like speech and music. At the same time, speech scientists, in conjunction with psycholinguists, have been developing models that attempt to derive the phonetic representation from the auditory data stream rather than taking it as given. As a result, there is now considerable interest in establishing common representations and determining where and how the reduction from a high-data-rate parallel system to a low-data-rate serial system occurs. Part 3 of die chapter outlines three approaches currently used to reduce the spectral representation of speech to phonology or words, and considers how each might be expanded to accommodate the high data rates flowing from spectro-temporal auditory models.

B. The Spectrogram and the Auditory Filter Bank

Prior to about 1950, hearing and speech were more closely related sciences in the sense that researchers who worked on one very often worked on the other. Much of the work on hearing was done with the explicit intention of developing a better understanding of speech perception and both groups took a basically psychological approach. About this time, however, many psychoacousticians turned away from speech and began to try and relate auditory perception to more peripheral, rather than more central, processes. Using linear systems analysis and signal detection theory, they built spectral models of masking and discrimination that related human perception to the frequency analysis performed by the basilar membrane. For simplicity, the models tended to concentrate on the peripheral activity produced by stationary sinusoids presented on their own or in noise.

Unfortunately, such models are of limited use to speech scientists trying to determine the critical auditory features required to distinguish, say, [e] from [a]. There were also practical constraints on the amount of computation that could be allocated to the front-end processor. For these and other reasons, many speech groups chose, effectively, to finesse the problem of auditory analysis by assuming that the spectrogram would serve as a sufficient front-end processor for speech stimuli.

1. The Spectrogram

A spectrogram of the word "past" spoken by an English Canadian is shown in Figure la; the vertical and horizontal dimensions are frequency and time, respectively. The central section of the figure with the vertical striations represents the vowel [ae]. vowels are voiced sounds, that is, they are quasi-periodic, and it is this property that generates the temporal regularity in the spectrogram. In contrast, there is an irregular patch of high frequency energy just after the vowel, which represents the [s]. It is an unvoiced speech sound (a burst of noise) and so there is a lack of temporal regularity in this region of the spectrogram. The dark horizontal bands in the vowel show concentrations of energy known as formants. In ASR, the position and trajectory of the formants are used to identify vowels. Recognition machines based on this kind of representation have had considerable success. They have the advantage of being relatively inexpensive and some of them operate in real time. As noted earlier however, there remains considerable room for improvement as performance is poor in noisy or reverberant environments.

The primary problem with the spectrogram is that it simply does not have sufficient resolution. It enables one to detect the presence of formants and to track their motion, but it does not have the resolution required to reveal the shapes of the formants within the pitch period, information that might be expected to assist with speaker identification and speaker adaptation. As a result it is not a satisfactory substitute for auditory analysis.

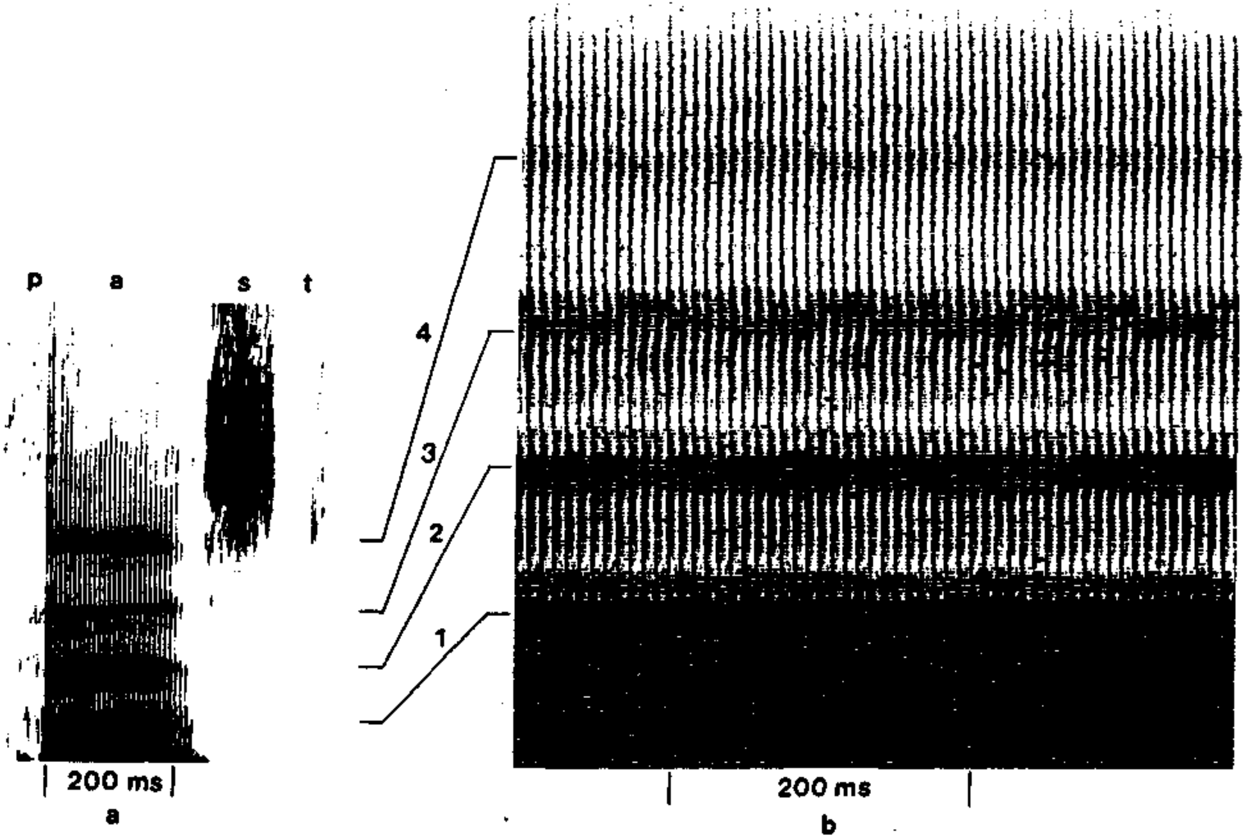

FIG. 3.1 Spectrograms of (a) the word "past" and (b) an enlargement of a sustained version of the vowel in "past". The abscissa is time, the ordinate is linear frequency, and the enlargement factor is 2.4. Note that the graininess of the enlargement is due to the resolution of the original spectrogram, and that the shapes of the formants are not apparent in this representation.

An enlargement of a sustained section of the vowel [ae] is presented in Figure lb to show that the resolution problem is not simply a matter of the overall size of the spectrogram. The blurry edges of the smallest features show that we have reached the limits of the resolution of this analysis method. The enlargement shows mat the energy in the formant track is not evenly distributed throughout the pitch period, and this indicates that more formant information is available at higher levels of magnification, but the shape of the formant within the pitch period does not exist in this representation.

It is also the case that the spectrogram is incompatible with both psychoacoustic and physiological representations of the auditory periphery, and so its use in speech research had the effect of increasing the gap between the speech and hearing communities in the 1950s and 1960s. With regard to psychoacoustics, the problem is that the spectrogram is a very poor predictor of auditory masking. The primary determinant of auditory masking is the bandwidth of the auditory filter which in normal adults increases from around 70 Hz at the low end of the speech range to around 700 Hz at the high end of the speech range. The spectrogram is like an auditory filterbank in which all of the filters have the same bandwidth. In the standard spectrogram, like that of Figure 1, the filter has a bandwidth of 300 Hz. As a result, the spectrogram over-estimates auditory masking at low frequencies and under-estimates it at high frequencies, to a degree that is simply unacceptable to psychoacousticians. With regard to auditory physiology, the problem with the spectrogram is that it integrates over too long a time, and so smears out the details of basilar membrane motion. As a result, it precludes any physiological model involving phase locking and any attempt to develop a realistic model of the firing patterns observed in the auditory nerve. Thus, the spectrogram is completely unacceptable to auditory physiologists as a representation of peripheral spectral analysis.

2. The Auditory Filter Bank

The separation between hearing and speech research persisted until about ten years ago, at which point the availability of more powerful computers made it possible to consider assembling full scale simulations of peripheral auditory processing (Young and Sachs, 1979; Dolmazon, 1982; Delgutte, 1980). At about the same time, psychoacousticians began to come to grips with the problems posed for their models by complex sounds (Yost and Watson, 1987), and speech scientists became concerned with the fidelity of their representations of speech sounds (Lyon, 1984; Schofield, 1985; Seneff, 1984). The net result is that there is a new common ground for hearing and speech research in the form of elaborate spectro-temporal auditory models, whose purpose is to characterise the patterns of information produced by complex sounds in the auditory nerve, and to process the patterns into a stream of auditory features and speech phonology (Beet, Moore and Tomlinson, 1986; Cooke, 1986; Gardner and Uppal, 1986; Ghitza, 1986; Hunt&Lefebvre, 1987; Patterson, 1987a; Shamma, 1986). The first stage in a spectro-temporal model is the auditory filter bank which perfonns the spectro-temporal equivalent of the spectral analysis that appears in the spectrogram. Figure 2 shows me output of a typical auditory filter bank when the input is the central portion of the [ae] in "past". As in Figure 1, the ordinate and abscissa are frequency and time, respectively, but in Figure 2 the time scale is greatly expanded. Whereas in Figure lb, the vowel occupies about half of the figure width and contains over 40 pitch periods, in Figure 2, the vowel occupies the entire width of the figure and contains only four pitch periods. Each of the fine lines in Figure 2 shows the output of a single auditory filter as a plot of amplitude versus time. There are 189 channels in this filter bank and the surface that the filter outputs define is intended to represent the motion of the basilar membrane. The filter bank is described in greater detail in Part 1.A. The important point here is to observe the overall patterns of motion produced by vowels.

The set of three features that occur in the upper half of each cycle of the vowel are the second, third and fourth formants of [ae] as they appear within the pitch period. In

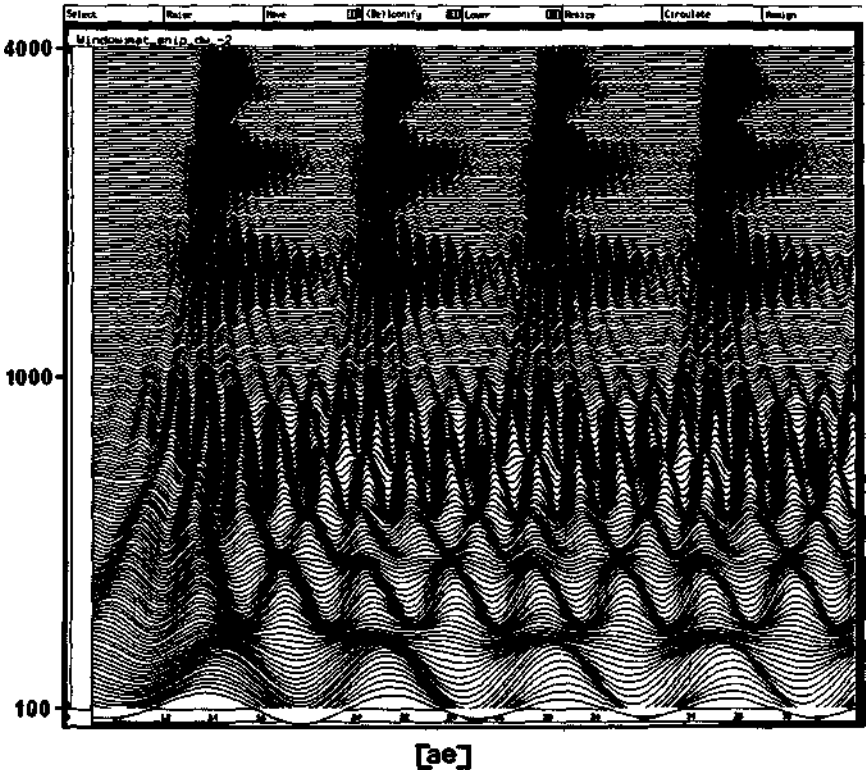

FIG. 3.2 A cochleogram of four cycles of the [ae] in "pasf produced by a gammatone auditory filterbank with 189 channels. The triangular objects are the formants. This representation shows that they have a distinctive shape that is not revealed in the spectrogram. The abscissa is time and the duration of each period is 8ms (fo =125 Hz). The ordinate is filter centre-frequency on an ERB-rate scale.

two dimensions (frequency and time), the formants appear as triangular objects whose temporal extent decreases as formant number increases. When we include the third dimension (filter amplitude), the shape becomes that of a cone with its core parallel to the tune axis. The interpretation of the first formant is more complex because, in that case, the temporal extent of the cone is greater than the pitch period and so the cones overlap and interact Nevertheless, it is also, basically, a cone. Thus, from the auditory perspective, the basic pattern of a vowel is a set of four regularly recurring, temporally coordinated cones. This set of physical characteristics is probably sufficient to identify a stream of sounds as speech rather than some other pitch producing event like music.

The patterns of motion produced by four different vowels, [i], [ae], [a] and [u], are shown in Figure 3. All four vowels are from the same speaker and the [ae] in the second panel is from the same vowel as that in Figure 2. In order to maintain the same scale as Figure 2, each vowel is restricted to one pitch period and in each case the period was selected from the centre of the vowel. At the bottom of each panel in Figure 3, one can observe a single cycle of the fundamental of the vowel, and just above it, two cycles of the second harmonic. Both show the sinusoidal motion characteristic of resolved, or isolated, harmonics.

All of the formants of [ae] occupy separate regions of the spectrum and the first

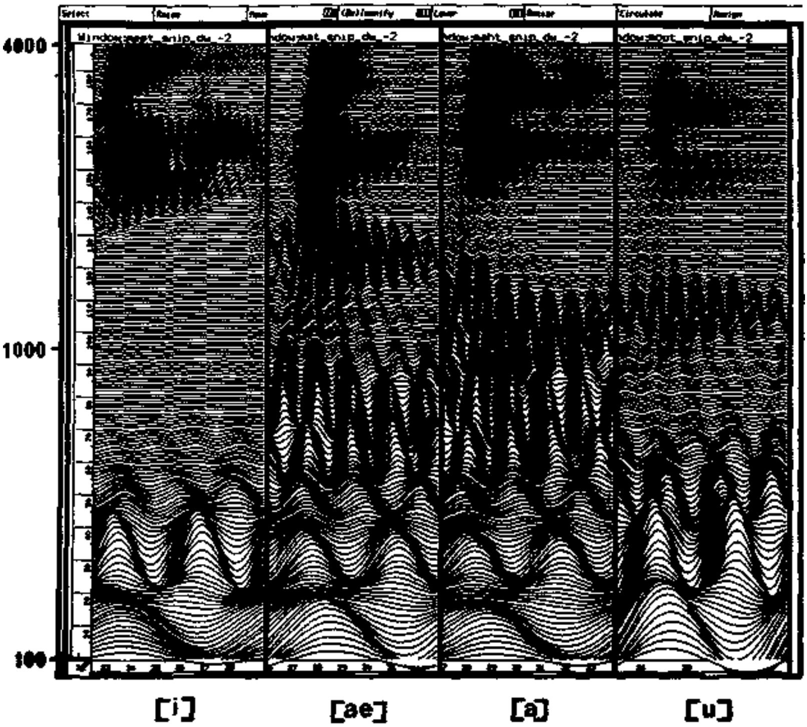

FIG. 3.3 Cochleograms of the four vowels [i], [ae], [a] and [u]. The position and strength of the formants identify the vowel.

formant is centred on the fourth and fifth harmonics. The first formant of [i] (leftmost panel) moves down from the position it occupied in [ae] into the region of the second harmonic, and the second formant moves up to encroach on the region of die third formant The first formant of [a] (third panel) occupies the same region as it does in [ae] while the second formant moves down into the region adjacent to the first formant. Both the first and second formants of [u] move down relative to their positions in [ae]. In this case, however, the more striking change is the reduction in the amplitude of the second, third and fourth formants. Taken together these observations suggest that a general purpose speech machine would benefit from the inclusion of a feature extractor that, in one way or another, fitted a set of four cones to the pattern of motion in each pitch period, and then used the summary values concerning the positions and sizes of the cones to identify vowels. The temporal information in the taper of the cones should provide for much more accurate formant positioning and tracking than is possible from a simple spectral representation.

The temporal information also has other uses. For example, in the [a] (third panel), there is an extension to the end of the fourth formant which probably represents an irregularity in the speaker's glottal waveform. The same feature appears in the third and fourth formants of [i] and there is a hint of it in the fourth formant of both [ae] and [u]. Temporal features of this form could be useful in speaker identification or speaker authentication systems. Note that the feature would be integrated out in a purely spectral representation of speech. Other potential advantages of spectre-temporal models will be presented in Part 1. It is sufficient to note at this point that there is reason to believe that die extra temporal information in auditory models will enhance the capabilities of ASR machines when our models and computers expand to the point where we can cope with die higher data rates.

1. PERIPHERAL AUDITORY PROCESSING

In the cochlea, there are four rows of hair cells along the edge of the basilar membrane. The hair cells in conjunction with me primary auditory neurons convert the motion of the basilar membrane into a complex neural firing pattern that flows from the cochlea up the auditory nerve to the auditory cortex. There is now a reasonable degree of consensus concerning the major characteristics of the electro-mechanical operations performed by the cochlea, that is, the auditory filtering process and neural transduction process. We begin this part of the paper by describing a typical cochlea simulation that illustrates the important characteristics of, and recent advances in, cochlear processing.

The operations performed by the cochlea are often presented as if they were the only processing performed by the auditory system prior to speech recognition. In fact there are at least four operations, or groups of operations, that are applied to the neural firing pattern after it leaves the cochlea and before it reaches the speech recognition system, and each plays an important role in conditioning the signal. In the latter half of this section we outline three of the operations and attempt to put them in perspective with regard to cochlear processing. The remaining operation, localisation, is omitted for brevity.

A. Cochlear Processing

1. Auditory Filtering

The classic early work by von Bekesy (1960) suggested that the action of the basilar membrane was like that of a lowpass filter. In contrast, psychophysical experiments of the same era (Wegel & Lane, 1924) showed that, at moderate levels at least, the frequency selectivity of the human auditory system was better characterised by a bandpass filter function. The discrepancy between the basilar membrane data and the psychophysical data eventually led to the assumption that there must be a neural filtering mechanism in the auditory system beyond the cochlea, and that the performance of normal listeners was the result of a pair of cascaded filters, the first electro-mechanical, and the second neural (Houtgast, 1974).

Bekesy's experiments were performed on cadavers and the signals were presented at extremely high intensities. Over the past decade, advances in the Mossbauer technique have made it possible to measure the motion of the basilar membrane at ever lower intensities. As the results came in, it immediately became clear that, at all but the highest levels, the basilar membrane provides bandpass rather than lowpass filtering. The lowpass form presented by Bekesy was an artifact of the extreme signal levels required by the techniques available to him at the time.

The new data caused a revolution in our conception of the cochlea. It is now assumed that there is an active mechanism that sharpens the low-frequency skirt of the filter before neural transduction. Subsequent investigation has shown that there is surprisingly good agreement between the new physiological data and that summarised in psychophysical models of the human auditory filterbank (Schofield, 1985). Together these findings indicate that we can eliminate the neural sharpening stage in our models of the peripheral auditory system (de Boer, 1983), a simplification whose importance is difficult to overestimate. It would appear to indicate that the relatively simple auditory filterbanks used in most psychological models do provide a reasonable representation of cochlear filtering.

The Gammatone Filterbank: The operation of a typical filterbank is illustrated in Figure 4 with the aid of a pulse train with a repetition rate of 125 Hz shown in Figure 4a. The filterbank contains 94 channels with centre frequencies ranging from 100 to 4,000 Hz and there are four filters per critical band. Each line in panel (b) shows the output of one filter when the stimulus is the pulse train in panel (a). The general equation for the filter shape is given by the gammatone function which was originally used by physiologists (de Boer & de Jongh, 1978; Johannesma, 1972) to describe the filter responses they obtained in single unit studies with cats. The equation for the filter is:

$$gt(t) = \quad t^{n-1} \quad \exp(-2(pi)bt) \quad \cos(2(pi)f_c t) \qquad (1)$$

where t is time, f_c is the filter centre frequency, n is the filter order and b is a bandwidth
pi = 3,141592..

parameter. The term gammatone refers to the shape of the impulse response of the filter. The first two terms of the equation are the familiar gamma distribution from statistics and they define the envelope of the impulse response. The cosine term is a tone when the frequency is in the auditory range, and it provides the fine structure of the impulse response.

Patterson and Moore (1986) have reviewed the data on the shape of the human auditory filter and shown that the Roex filter shape suggested by Patterson, Nimmo-Smith, Weber and Milroy (1982) provides a good approximation to the human filter shape over a wide range of stimulus conditions. Recently, Schofield (198S) has shown that die gammatone function can provide a good fit to the human filter-shape data measured by Patterson (1976), indicating that the gammatone filter and the Roex filter are close relatives. The gammatone filter has the advantage of providing both a

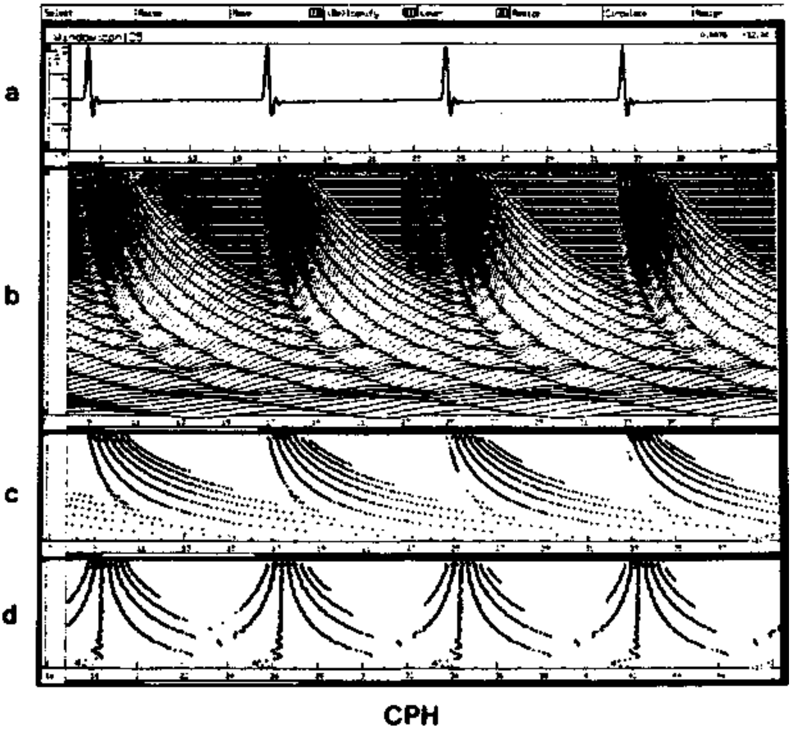

FIG. 3.4 The processing of a pulse train, or CPH wave, by the pulse ribbon model. The filterbank converts the wave (a) into a cochleogram (b) which the array of hair-cell simulators convert into a pulse ribbon, either without (c), or with (d), phase compensation.

spectral and a temporal representation of the filtering process. Accordingly a gammatone filterbank with parameter values that represent those found in human hearing has been developed, and it is this filterbank mat underlies the illustrations in this part of the paper.

The gammatone expression was tuned to human hearing by (a) setting the filter order, n, to 4, (b) distributing the filters across frequency as suggested by Moore and Glasberg (1983), and (c) calculating the parameter b using the ERB function:

$$ERB(fc) = 6.23 \text{ x } 10^{-6}\, fc^{2} + 93.39 \text{ x } 10^{-3}\, f_{c} + 28.52 \quad (2)$$

and the scaling relationship

$$b = 1.019\ ERB(fc). \qquad (3)$$

Each filter is then convolved with the signal to produce one of the channels of output in panel (b) of Figure 4. The surface defined by the array of outputs represents the motion of the basilar membrane. The individual filter outputs are referred to as driving waves because they "drive" the hair cells in the sense of determining the temporal pattern of the spikes in the pulse streams that flow up the auditory nerve.

The output of the filterbank is quite different from that of a magnified spectrogram like that shown in Figure lb because the bandwidth of the filter increases with centre frequency in the auditory filterbank. The driving waves in the lower part of Figure 4b are from relatively narrow filters centred in the region of the first four harmonics of the pulse train, and they are essentially sinusoidal in shape. Those in the middle part of the panel are from wider filters centred near harmonics 5 to 12, and they are more like amplitude-modulated sinusoids. The "carrier" frequency is approximately the centre frequency of the filter and the "modulation" frequency is the repetition rate of the pulse train. The modulation depth increases with centre frequency as the filter broadens and the attenuation of adjacent harmonics decreases. The driving waves in the upper part of panel (b) are from relatively wide filters centred near harmonics 13 to 32. In this region, the outputs are like a stream of individual impulse responses because the integration time of the filter is short with respect to the repetition rate of the pulse train. In a system with proportional bandwidth, the pattern of membrane motion is relatively independent of the repetition rate of the stimulus; the cycles move closer together as the pitch rises and the energy associated with individual harmonics moves up the figure somewhat, but the pattern remains largely unchanged.

The pronounced rightwards skew in the lower half of the filterbank output is also caused by the fact that filter bandwidth increases with centre frequency. But there is evidence from phase perception studies that the auditory system compensates for the phase lag that produces the skew (Patterson, 1987b). As a result, we often apply a phase compensation to the cochleogram in order to bring together vertically those parts of the pattern that belong to one pitch period of the original sound. The vowel cochleograms presented in the Introduction are a case in point They were generated by a gammatone filterbank and phase-compensated to align the formants.

2. Neural Transduction

The motion of the basilar membrane is converted into nerve impulses by the hair cells and the primary auditory neurones of the eighth nerve. Physiological research over the past two decades has revealed several important facts about neural transduction:

1) The hair cell applies something like logarithmic compression to the amplitude of the driving wave.
2) The adaptation we observe in the auditory nerve takes place in the hair cell and the synaptic cleft that separates it from the primary neurone that it drives.
3) There are few cross connections in this part of the system; by and large, the outer hair cells amplify membrane motion for the inner hair cell, which in turn drives the primary neurone.

These advances have led the physiologists to suggest relatively simple, "reservoir" models of neural transduction (Schwid and Geisler, 1982; Meddis, 1986). In practical terms, it would appear that a reasonable approximation is provided by a device consisting of a logarithmic compressor followed by a peak picker that has one fast and one slow adaptation parameter. Such a unit produces a phase-locked stream of pulses that preserves information concerning the times between die positive peaks in the wave, like the streams observed in auditory nerve fibers.

The Initial Pulse Ribbon: The cochlea simulation uses the hair-cell simulation suggested by Meddis (1986). A bank of "hair cells" converts the 96 driving waves into 96 pulse streams as illustrated in Fig. 4c. Each pulse stream is intended to represent the output of all the fibers associated with one frequency channel. In short, the stochastic properties of neural transduction are ignored for the moment, and a volley mechanism of some sort is assumed. In this case, sinusoidal driving waves like those in the lower portion of panel (b) are converted into regular pulse streams with one pulse per cycle as shown at the bottom of panel (c). Modulated driving waves like those at the top of panel (b) are converted into modulated pulse streams in which bursts of pulses are regularly separated by gaps as shown at the top of panel (c). The period of the carrier frequency is equal to the time between pulses within a burst and the period of the modulation frequency is equal to the time between corresponding pulses in successive bursts.

Collectively, the array of pulse streams is referred to as the "initial pulse ribbon" and it provides an overview of the information flowing up the auditory system from the cochlea. The horizontal dimension of the ribbon is "time since the sound reached the eardrum"; the vertical dimension is "auditory-filter centre frequency" which is a roughly logarithmic frequency scale. If the brightness of each channel were varied to reflect its current amplitude, the initial pulse ribbon would be like a spectrogram with an expanded time scale. For a periodic sound, the pattern repeats on the ribbon and the rate of repetition corresponds to the pitch of the sound. $ltimbreTimbre corresponds to the pattern of pulses within the cycle. The pattern has a spectral dimension (vertical) as in traditional spectral models, but it also has a temporal dimension (horizontal), and the fine-grain information on the latter dimension enables the ribbon to represent phase-related timbre changes.

The initial pulse ribbon, then, is a device for presenting the temporal information and the phase information of the auditory nerve, in a form where we can better appreciate the patterns of information generated by complex sounds like music and speech. It is not intended to be new or controversial but rather to provide a simplified view of what comes out of the cochlea to support further research.

The bottom panel of Figure 4 shows the initial pulse ribbon produced by the pulse train when phase compensation is included in the operations. The compensation brings together in a vertical column those pulses associated with the largest peaks in the cochleogram, and it helps to emphasise the natural symmetry of this stimulus. For comparison, the phase-compensated pulse ribbon produced by the [ae] of Figure 2 is presented in Figure 5. The largest cochleogram peaks are also aligned in this figure, but the formants impose a spectro-temporal weighting that imparts a strong left/right

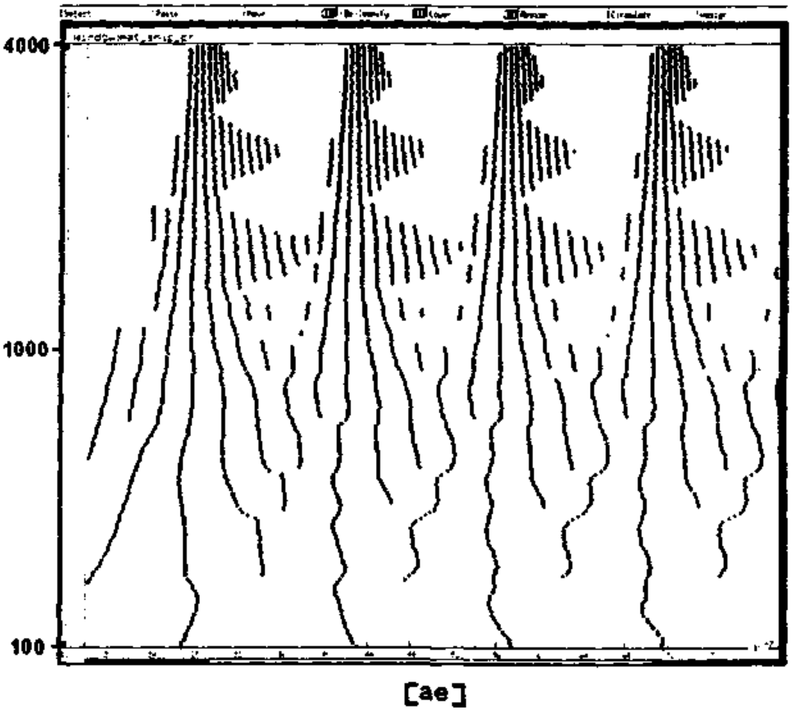

FIG. 3.5 The initial pulse ribbon produced by four cycles of the [ae] in "past". Note that this reduced representation preserves the basic pattern of information in the cochleogram of Figure 2.

asymmetry, an asymmetry which is characteristic of voiced speech sounds. Note also that the pulse ribbon preserves the basic information of the corresponding cochleogram even though it requires less than one tenth the bandwidth.

B. Neural Peripheral Processing

There are now a number of physiological and psychological models of hearing that include some representation of auditory neural processing as well as cochlear processing. It is still the case, however, that physiological models tend to emphasize cochlear processing and include only the earliest stages of the neural processing. As a result, they are usually less appropriate as preprocessors for ASR than psychological models which combine functional models of cochlear processing with functional models of more central processes, such as pitch perception. There is not space in a chapter this size to compare physiological and psychological models of hearing with regard to their suitability as ASR preprocessors. Rather we will present one psychological model which makes an explicit attempt to be comprehensive and to balance the level of complexity used in the representations of cochlear and neural processing.

The "pulse ribbon" model of hearing (Patterson, 1987a) was originally created to provide a bridge between the output of the cochlea as observed in single nerve fibres of small mammals stimulated by simple sounds, and the sensations that humans hear when stimulated by complex sounds like music and speech. The model has five stages: the first two stages simulate auditory filtering and neural transduction and they form the cochlea simulation just described in Section A. With regard to cochlear processing, the pulse ribbon model is like most other psychological models, attempting to summarise our knowledge concerning frequency selectivity and neural transduction in the form of an array of pulse streams.

The remaining three stages transform this initial pulse ribbon using operations that are intended to characterise phase perception, pitch perception and timbre perception, respectively. Together they illustrate the kind of neural processing required to convert the output of the cochlea into stabilised patterns that represent the perceptions, or auditory images, produced by sounds. In the model these stabilised pulse patterns are the output of the peripheral auditory system and the input to more central systems like those for speech and music.

1. Phase Perception

In 1947 Mathes and Miller proved that, contrary to previous suggestions (Helmholtz, 1875,1912), the auditory system is not phase deaf. They showed mat changes in the envelopes of high-frequency driving waves change the timbre of the sound. Confirmations and extentions of their findings have been reported at regular intervals since that time (for a review see Patterson, 1987b). For over 50 years, then, throughout the development of spectral front-ends for speech processing, we have known that strictly spectral models of auditory processing must ultimately fail, and that, at best, these models are a practical simplification of peripheral processing.

With hindsight, there are two obvious reasons for ignoring the data on phase sensitivity, over and above the fact that they would have rendered models unacceptably complex at that time: firstly, the timbre changes produced by phase changes were not thought important for speech perception and secondly, there were no coherent models of phase perception to unify the observations and suggest how it might be implemented. Recent research, however, has changed both of these positions. With regard to the first, it is now clear that phase changes produce relatively strong perceptual effects (Carlson et al, 1980), and that they almost undoubtedly do play a role in vowel discrimination (Tranmuller, 1988). Furthermore, there is the hint that it is the proper handling of phase information that enables the auditory system to perform so much better than speech recognisers in reverberant environments. With regard to the second, coherent models of phase perception are beginning to appear (Patterson, 1987b; Wakefield, 1987).

This subsection describes a series of phase experiments to illustrate the advances that have been made in our understanding. The experiments are from Patterson (1987b) and they were performed to determine our sensitivity to changes in the envelopes of the driving waves, changes introduced by local alterations of the phase spectrum. In each case, the stimuli were composed of 31 equal-amplitude harmonics of a fundamental, fo, and all that varied was the phase spectrum of the stimulus. The stimuli were "alternating-phase" waves in which all of the odd harmonics were in cosine phase while all of the even harmonics were in some other fixed phase, D. Figure 6a shows the alternating-phase wave when D is 40 degrees. When D is 0 the wave is a pulse train, or cosine-phase wave. As D increases the secondary peak in the middle of the cycle grows and eventually we hear a timbre change. In the mid- to high-frequency channels of the filterbank, the secondary peak in the sound wave causes a local maximum in the envelope of the driving wave midway between the main envelope peaks (panel b) and the size of the local maximum increases with D. When D is large, the local maxima cause the pulse stream generators to produce an extra column of pulses in the initial pulse ribbon (compare Figures 4c and 6c) and it is these pulses that are assumed to produce the timbre change. The alternating-phase stimulus was used to map out the existence region for local phase changes.

The wave in Fig. 6a is just discriminable from a cosine-phase wave when the fundamental is 125 Hz and the level is 45 dB/component. When 0 is lowered by an octave, the period of the wave doubles. In this case, the pulse generators have effectively twice as many pulses to assign to each cycle of the driving wave and the local maxima appear in the pulse ribbon at a lower D value. Thus, the model predicts that timbre threshold will be strongly affected by the pitch of the stimulus, and this is indeed the case. The firing rates of auditory nerve fibers increase with stimulus level which suggests that the sustained firing rates of the pulse generators in the model should vary with stimulus level. Increasing the model rates causes the local maxima in the driving waves to appear in the pulse ribbon at a lower D value and so the model predicts that timbre threshold will vary inversely with stimulus level, and this prediction is also borne out by the data. Thus, it would appear that a pulse ribbon model can account for the timbre changes associated with envelope changes in terms of the firing rates of the

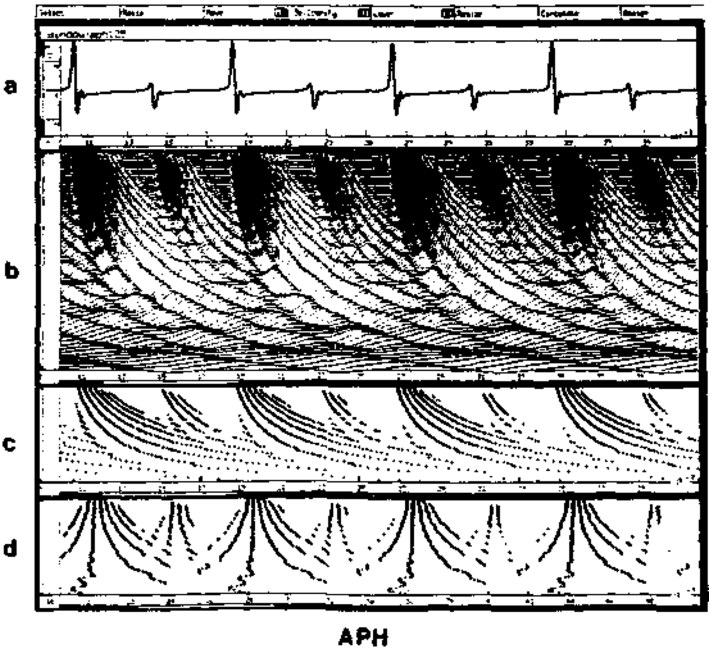

FIG. 3.6 The processing of an APH wave by the pulse ribbon model (D=40 degrees). The filterbank converts the wave (a) into a cochleogram (b) which the array of hair-cell simulators convert into a pulse ribbon, either without (c), or with (d), phase compensation. Note that the secondary pulse in the waveform produces a secondary ridge in each cycle of the cochleogram. The resulting feature in the pulse ribbon is assumed to be the cue that mediates the timbre change associated with this stimulus.

pulse stream generators, and conversely, the data from alternating-phase experiments can be used to set the parameter values in the model.

2. Pitch Extraction

The purpose of the fourth stage of the model is to determine the pitch of the sound. Originally, speech recognition devices extracted the pitch value from a lowpass filtered version of the acoustic waveform. Although this works reasonably well when the speech occurs in a quiet environment, it fails when the speech occurs in a noisy environment. More recently, speech and hearing models have come to use pitch extractors based on one of the "central spectrum" models of pitch perception (Wightman, 1973; Goldstein 1973; Terhardt, 1974). In their original forms, these models ignored the timing information in the driving waves and estimated the pitch solely on the basis of the power, or overall firing rate, in each channel. In essence, it was argued that the overall-rate information was sufficient to explain the psychophysical data so long as optimal use was made of that information. These pitch extractors operate better in noise than those that operate directly on the acoustic waveform, but they are still not all that good. Furthermore, devices that discard the fine-grain temporal information do not perform well when the peak of the glottal pulse is degraded as in reverberant rooms.

For these and several other reasons, a number of groups have chosen to investigate the potential of spectro-temporal models which, as the name suggests, use the temporal, as well as the spectral, information (Young and Sachs, 1979; Goldstein and Srulovicz, 1977; Lyon, 1984; Gardner and Uppal, 1986; Cooke, 1986; Beet, Moore and Tomlinson, 1986; Patterson, 1987a). At the output of the cochlea, a periodic sound produces a repeating pulse pattern (Figures 4d and 6d) and the repetition rate of the pattern provides a good estimate of the pitch of the sound. The voiced parts of speech are quasi-periodic sounds and they also produce repeating pulse ribbons as illustrated in Figure 5. Spectre-temporal models make use of the spectral information in the sense that they separate the signal energy into different frequency bands. In addition, however, there is a second frequency analysis, — a temporal analysis, performed neurally, mat extracts the repetition rate of the pattern flowing up the auditory system.

The Spired Processor: One attempt to solve the problem of temporal frequency analysis is the "spiral processor" suggested by Patterson (1986, 1987a). Briefly, the temporal regularity observed in the pulse ribbons of periodic sounds can be converted into position information if the pulse ribbon is wrapped into a logarithmic spiral, base two. For example, consider the pulse ribbon associated with the alternative-phasing wave (Figure 6d). If we assume that the temporal window on which the periodicity mechanism operates is 72 ms in duration, then it will contain 9 cycles of the pulse ribbon at any one moment as shown in the upper panel of Figure 7. If this pulse ribbon is wrapped into a spiral, base 2, the result is the 9-cycle spiral ribbon shown in Figure 7b. The threads of the pulse ribbon in panel (a) become a set of concentric spirals in panel (b). The outer and inner strands of the spiral ribbon contain the pulse streams from the 1st and 96th channels, respectively.

The pulses appear at the centre of the spiral as they are generated and flow along the spiral as time progresses, dropping off at the outer end 72 ms after appearing. So time itself keeps track of the pulses as they are being correlated with their neighbours in time and space. The stimulus occupies four revolutions of the spiral, and at the moment shown, four of the vertical columns that mark cycles on the pulse ribbon are themselves lined up on one spoke of the spiral, the vertical spoke emanating from the centre of the spiral. A unit monitoring this spoke would note above average activity at this instant and so serve as a detector for 125 Hz. A stable display of the current pitch pattern can be obtained from the continuously flowing spiral ribbon, by strobing die display when the pulse coincidence occurs. The angles between the spokes are the same no matter what the note; it is only the orientation of the spoke pattern mat changes when the pitch is altered. As the pitch rises, the spokes rotate clockwise as a unit and the pattern completes one full revolution as the pitch rises an octave. Computationally, the spiral processor is just a mapping that warps the space through which the pulses

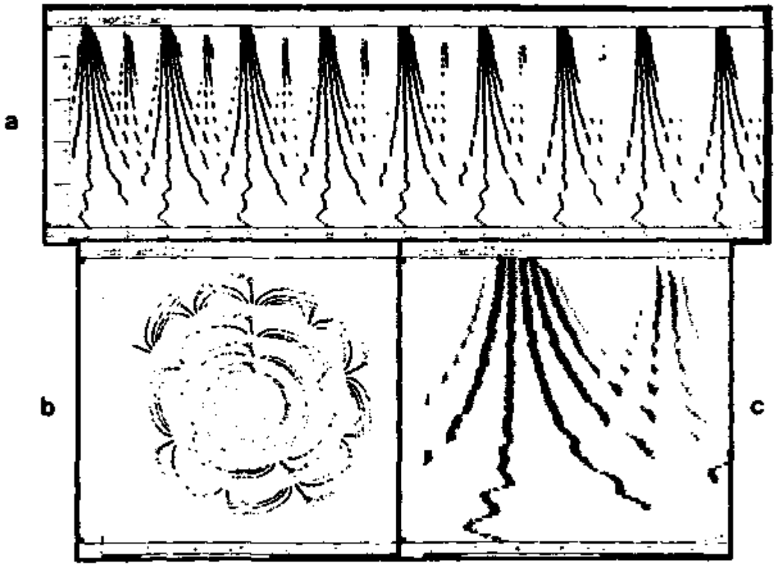

FIG. 3.7 Pitch extraction and timbre stabilisation in the pulse ribbon model. The phase compensated pulse ribbon (a) is wrapped into a logarithmic spiral (b) to extract the pitch, and wrapped into a cylindrical pulse ribbon (c), with circumference equal to the pitch period, to stabilise the repeating pulse pattern.

flow, so that clusters of pulses that repeat in time come together in space for an instant and produce a secondary pulse indicating the pitch of the sound. Since it is a mapping it can be implemented as a table look-up operation, which makes it a relatively efficient process.

3. Timbre Stabilisation

The pulse patterns produced by successive cycles of a periodic wave are highly correlated. The timbre of the sound is coded in these pulse patterns and so one should combine them to obtain the best estimate of timbre in the statistical sense. In the pulse ribbon model this is accomplished, once the pitch of the source is known, by wrapping the pulse ribbon around a cylinder whose circumference is the period of the original sound. In this case, successive cycles of the ribbon fall on top of each other and form a stabilised image of the timbre pattern for as long as the sound is stationary. When the input is noise, the pulse streams are not periodic and the timbre pattern is a rectangular random dot pattern no matter what the diameter of the cylinder.

The timbre pattern for the alternating-phase wave is shown in panel (c). For convenience, a planar display is used as if the cylindrical ribbon had been slit down the back and flattened out There are approximately 4000 pulses in the initial pulse ribbon of panel (a) and all of them are plotted in panel (c) as well. However, since the sound is periodic, successive cycles of pulses coincide and in this way the correlated timbre information is combined. The column of pulses on the righthand side of the timbre pattern gives the sound its distinctive timbre. Purely spectral models that only use the overall-rate of firing effectively integrate across the temporal dimension of the ribbon and obscure this feature. Although we pitch of complex sounds is singularly insensitive to the phase of their constituent components (Patterson, 1973; Patterson and Wightman, 1976), nevertheless, the timbre of complex sounds is affected by component phase (see Patterson, 1987b, for a review), and vowel discriminations are largely timbre discriminations. Thus, a stabilised timbre pattern should assist the extraction of those auditory features that indicate the presence of a speech sound.

C. Conclusion

1. Cochlear Processing

The primary conclusion with regard to the earliest stages of auditory processing is mat there now exist relatively simple simulations of auditory filtering and neural transduction that together provide a much better representation of cochlear processing than does the spectrogram. The replacement of spectrographic and similar place representations with a cochlea simulation should enhance ASR performance even if the remaining stages of auditory processing are ignored.

2. Neural Processing

Phase: The auditory system is phase-sensitive and the inclusion of a competent neural phase mechanism should improve ASR performance in the areas of speaker identification and resistance to reverberation.

Pitch: Until recently, pitch was not thought to be a particularly important variable in speech recognition. It was argued that although pitch is a major determinant of prosody, nevertheless, large changes in prosody do not prevent one recognising individual words in a phrase or sentence. The pulse ribbon model leads us to conclude that pitch is not just one of many speech features, it is the key feature that makes it possible to stabilise the timbre of the voiced parts of speech and so extract the remaining speech features more effectively. A similar concept already exists in speech research, where the use of pitch information to create a better vowel representation is referred to as "pitch-synchronous feature extraction" (Seneff, 1984). However, the technique is based on driving-wave envelopes rather than neural firing patterns, and this will probably lead to a timbre image that is not quite as well focused.

In speech, the pitch varies in the short term over a range of about an octave. But much of the time, the rate of change is relatively slow when measured in auditory terms. In this case, the pitch extractor, whatever its form, can track the pitch and feed the current value forward so that the timbre pattern can be continuously adjusted to maintain a stable image. The image will rise or fall a little on the timbre display but die variation will be small relative to the change in pitch, and the pattern will remain identifiable.

Timbre: It is now possible to create a stabilised auditory image of stationary sounds, and the concentration of timbre information on the cylindrical pulse ribbon should assist feature extraction and speech segmentation. Currently auditory front-ends send a frame of timbre information forward to the next module in the speech processor every n milliseconds, even if there is no sound coming in. If, instead, the auditory front-end were set to check whether the pattern had changed, and to send a frame forward only when there was a significant change in the pattern, it would greatly reduce the computational load on the recognition system.

2. RECOGNITION

A. Background

In Part 2 we switch our attention to psycholinguistic work on the understanding of spoken language. Of course this differs from psychoacoustic research primarily in the level of processing under consideration. But the flavour of the research is also very different. Whereas psychoacousticians know, for instance, the nature of the operations performed on auditory signals by the basilar membrane and the hair-cell array, and tie their theorising closely to the physical characteristics of these structures, psycholinguists can call on no such physical constraints. Psycholinguists are cognitive psychologists, and their conceptual repertoire is accordingly restricted to cognitive constructs. The most central of these in the present context is recognition, i.e. acknowledgement that an input has been previously encountered. Obviously the concept of storage in memory is central to recognition, and so is the notion of a representation, or code in terms of which an input and a stored form can be matched with one another so that recognition can be achieved. In the following sections we discuss the basic characteristics of the recognition task as seen by psycholinguists, and die assumptions which underlie psycholinguistic research on the processing of spoken language. As we stated above, we examine mis processing only as far as the point at which spoken word recognition has been accomplished. Although there is a substantial body of psycholinguistic research on higher levels of processing, we will omit it entirely, for several reasons: it would extend the present discussion out of proportion to the rest of this chapter, it is discussed elsewhere in this book; and it makes no separate contribution to the problem of interfacing psychoacoustic work on auditory perception with psycholinguistic work on speech recognition.

B. Nature of the Recognition Task

Recognition involves matching an input to a pre-stored representation. In the case of speech recognition the input is an auditory representation and the pre-stored representation is conceptual; so speech recognition consists in the translation of sound to meaning. The goal of the task is achieving an internal representation in the recogniser that is equivalent to an internal representation in a communicator-recognition of the communicator's "message". The cognitive system takes as input a representation which is the output of auditory preprocessing, and it outputs in turn a selection from its stored set of sound-meaning associates.

Precisely how it does this is in part determined by the characteristics of these stored meaning representations themselves. The set of potential messages is infinite. But recognisers do not have infinite storage capacity. Therefore the stored set of representations, which is usually termed the *lexicon,* cannot possibly include every message a recogniser might potentially encounter. The set of representations in the lexicon must *be finite,* and it must consist of *discrete* units.

Part of the process of translating sound into meaning, therefore, must consist in determining which parts of a signal correspond to which discrete stored units. This is essentially a *segmentation* problem. Logically, the only segmentation of a speech signal which is required is segmentation into lexical units; as we shall see below, however, other segmentation units may be warranted in practice.

C. The Lexicon

Several characteristics of the lexical store are relevant to consideration of the segmentation issue. Firstly, the size of the discrete units represented in the lexicon must be highly variable. It is reasonable to suppose that many orthographically defined words will merit a separate stored representation, though of course there is no reason to suggest that it is a necessary criterion that each lexical representation be a separate orthographic word. Nor is it by itself a sufficient criterion, since orthographic words exist which have no separate conceptual representation (e.g. *kith,* which occurs now only in *kith and kin);* grammatical words *(to, the, but* etc.), whose "meaning" is their function in context, similarly present difficulties of conceptual definition. Some studies of the mental lexicon (e.g. Friederici & Schoenle, 1980) have proposed mat grammatical words are represented separately and differently from the greater part of the lexical stock. Similarly, it has been suggested (e.g. by Taft, 1988) tiiat affixal or stem morphemes may be stored as separate units (in English, for example, this would mean such separate entries as *un-, re-, -mit, -vert, -ment, -ish* etc.; but in highly affixing languages such as Turkish the set of potential stored morphemes would be very large indeed). It has also been proposed tiiat certain stored units may contain sequences of words, forming, for instance, idioms such as *kick the bucket* (Swinney & Cutler, 1979) or highly frequent expressions such as *good morning* (van Lancker, 1975). Since even monomorphemic words can vary dramatically in length *(owe, salmagundi),* it is clear tiiat die stored representations in the lexicon will be highly heterogeneous. Aitchison (1987) reviews recent research on the structure of the mental lexicon.

Secondly, whatever the constitution of the stored set of representations, its size is sure to be very large. Estimates of the educated adult language user's vocabulary have proposed an average size of 150,000 words (Seashore & Eckerson, 1940). To search a set of this size at a thousand items per second would take several minutes. Yet this is hardly a realistic estimate of lexical access time for a human recogniser (nor is it an acceptable goal for a commercial automatic recogniser). Both the size and the heterogeneity of the lexicon have implications for prelexical aspects of the recognition process, as will be outlined below.

D. The Normalisation Issue

The speech signal corresponding to a particular lexical representation is not a fixed acoustic form. It is no exaggeration to say that even two productions of the same utterance by the same speaker speaking on the same occasion at the same rate will not be completely identical. But within-speaker variability is tiny compared to the enormous variability across speakers and across occasions. Speakers differ widely in the length and shape of their vocal tracts, as a function of age, sex and other physical characteristics; productions of a given sound by a large adult male and by a small child have little in common. Situation-specific variations include the speaker's current physiological state; the voice can change when the speaker is tired, for instance, or as a result of temporary changes in vocal tract shape such as a swollen or anaesthetised mouth, a pipe clenched between the teeth, or a mouthful of food. Other situational variables include distance between speaker and hearer, intervening barriers, and background noise. Yet acoustic signals which (for all these reasons) are very widely varying indeed are nevertheless perceived by listeners as the same speech event. For this to happen, there has to be some way of factoring out the speaker- and situation-specific contributions. This is called the problem of *normalisation* across speakers.

A further source of variability is due to different varieties, or dialects, of a given language. Sounds can be articulated very differently in different dialects (compare, for instance, English /r/ as spoken in Kansas and in Boston, in Bombay, Aberdeen, Sydney, Somerset and Surrey), dialects also differ in the size of their phonemic repertoire (Southern British English uses different vowels in each of *book, but* and *boot,* but Scottish English has the same vowel in *book* and *boot,* and a different vowel in *but,* while Northern British English has the same vowel in *book* and *but* but a different vowel in *boot.)* Thus listeners have to normalise for dialect variability as well. At the word level, variability also arises due to speech style, or register, and (often related to this) speech rate. Consider the two words "did you". In formal speech they would be pronounced [dldju]; a phonetic transcription shows five separate segments. A more casual style allows for the [d] and [j] to palatalise to an affricate [dZ], giving [dldZu]. If the two words occur at the beginning of a phrase, the entire first syllable will often be dropped, leaving only the affrication as a trace of the word "did": "[dZu] see that?" Finally, in appropriate contexts the vowel of "you" can be reduced or lost entirely: "[dZ@] see it?"; "[dZit] yet?" In that latter phrase the single affricate [dZ] is performing the function of [dldju] in a formal, precise utterance of "did you eat yet?"; but there is no segmental overlap between the two transcriptions.

At the phoneme level, variability is further complicated by the phenomenon of coarticulation. Phonetic segments may be spoken quite differently as a function of the other segments which surround them in a particular utterance. Stop consonants are particularly sensitive to the identity of the following vowel; thus spectrograms of the words "past" and "pieced" look quite different in the initial consonant as well as in the vowel portions. In some cases these differences can even be noticed by the speaker (/k/ is articulated further forward in the vocal tract in speaking "keen" than in speaking "corn"). Moreover, coarticulation effects are not confined to immediately adjacent segments; their influence can stretch both forwards and backwards over several segments. Consider the utterance "she has to spruce herself up"; in most cases, the lip-rounding for the [u] in "spruce" is fully in place by the utterance of the word-initial [s], or even during the preceding syllable; and it does not disappear until well into the word "herself'.

This extreme variability means, simply, that if the lexicon were to store an exact acoustic representation for every possible form in which a given lexical unit might be presented as a speech signal, it would need infinite storage capacity. Therefore the lexical representation of the input signal, i.e. the sound component of the sound-meaning pairing, must be in a relatively *abstract* (or normalised) form. In consequence, the progression from auditory featuresto the input representation for lexical access necessarily involves a process of *transformation*.

E. The Continuity Issue

Units of lexical representation (words) are all that it is necessary to locate in the input But the nature of auditory linguistic input is that it extends over time - a portion of input corresponding to a particular lexical form is not simultaneously available in its entirety. Moreover, only rarely are recognisers presented with isolated lexical items. Most speech signals are made up of a stream of words, and the stream is effectively continuous in that momentary discontinuities within it do not correspond systematically to its linguistic structure.

The importance of the continuity issue for speech recognition has been neglected, simply because the majority of psycholinguistic studies of lexical storage and retrieval have been carried out in the visual domain. In nearly all orthographies, representations of linguistic messages in the visual domain consist of discontinuous units: words, which are made up in turn (depending on the orthography) of letters, syllables or the like. Under such circumstances, segmentation is no problem. Whatever the orthography, explicit markers in the input (i.e. spaces) signify the boundaries of portions of the input corresponding to lexical units; each of these units may then be further subdivided into elements which offer a possible subclassification scheme for the lexicon and hence a possible route for efficient lexical access, segmentation in the auditory domain would be similarly unproblematic if explicit boundary markers signalled which parts of the signal belonged together in a single lexical unit. Years of research in speech science, however, have failed to isolate reliable cues to lexical boundaries. One way round this problem is simply to match arbitrary portions of the auditory input (subject, of course, to suitable transformation) against lexical templates. This crude process, in a number of different guises, is in fact the basis of all automatic speech recognition systems currently in commercial production. Such template-matching procedures are, however, extremely inefficient Firstly, they involve a large number of futile access attempts, since the heterogeneity of lexical units means that the duration of the string to be tested cannot be predicted. Secondly, since they invoke a simple search procedure, the large size of the lexical stock means that each attempt at access requires a long search. This is one reason why all current commercial automatic speech recognisers are limited to very small vocabularies.

The problem *of segmentation* under conditions of *continuity* suggests that prelexical classification of speech signals into some sub-word-level representation might enable a more efficient system of lexical access. As letters or syllables in orthography open up the possibility of classification within the lexicon and an access procedure based on this classification, so do sublexical units in the auditory domain. This overcomes the necessity for simple search procedures in lexical access, and hence removes the problem of the impracticable amount of time required to search a vocabulary of the size used by human recognisers. But the greatest advantage of a sublexical representation is that the set of potential units would be very much smaller than the set of units in the lexicon. However large and heterogeneous the lexical stock, any lexical item could be decomposed into a selection from a small and finite set of sublexical units. The normalisation issue and the consequent necessity of *transformation* provides another argument in favour of an intermediate level of representation between auditory featuresand lexical input. If transformation is necessary in any case, then transformation into a small set of possibilities will be far easier than transformation into a large set of possibilities.

F. Prelexical Representations

Psycholinguists have devoted a great deal of research effort to investigating the form that prelexical representations should take. For a segment of a speech signal to function as such a unit of representation, there are three conditions which it should meet:

1. The segments themselves, at whatever level they are, must be reasonably distinguishable in the speech signal. Note that this does not imply that they must have explicitly marked boundaries. If the boundaries of any sublexical unit were explicitly marked, then the boundaries of words would *ipso facto* be explicitly marked, but, as we have already observed, this is not the case.

2. The whole utterance must be characterisable as a string of the segments in question, with no parts of the utterance unaccounted for. (Thus although fricative noise might satisfy the first requirement, it is not acceptable to propose the interval from one fricative to the next as a potential sublexical unit of representation, since utterances may contain no fricatives at all.)

3. The units must correspond in some reliable way to lexical units. That is, if the unit in question is not necessarily sublexical, then some simple and predictable translation from the prelexical unit to the lexical unit should be possible.

Most current models of lexical storage and retrieval for spoken word recognition assume that (for the theoretical reasons outlined above) human recognition does involve some prelexical level of representation. It is assumed that this representation encodes the input in a form which can serve to access the lexicon efficiently, i.e. it corresponds to the code used on the "sound" side of the lexical sound-meaning pairings. In practice, the most obvious candidates for the role of intermediate representation have been the units of analysis used by linguistic science. The phoneme has been the most popular choice because (by definition) it is the smallest linguistic unit into which speech can be sequentially decomposed. The syllable is the second most popular choice; it is the smallest linguistic unit which can be independendy uttered (with the exception, admittedly, of those phonemes which are realised as hisses or buzzes).

A great deal is known about the nature and manner of use of acoustic cues for identifying and distinguishing between phonemes, from speech perception work within linguistics and phonetics; see Pisoni and Luce (1986) and Jusczyk (1986) for reviews of this work. The most central issue in this debate for decades has been the question of invariance (see Perkell & Klatt 1986), i.e. the degree to which acoustic cues to phonemes can be said to possess invariant properties which are necessarily present whenever the phoneme is uttered, and which are therefore resistant to the sources of variation described in section D above. Insofar as syllables are made up of phonemes, this work is equally relevant to the perception and identification of syllables.

But this body of research, which has been conducted principally by phoneticians, is to a certain extent orthogonal to the psychological question of whether either phonemes or syllables are a necessary or appropriate level of representation for lexical access from auditory input. The question at issue here is, chiefly, whether phonemes or syllables constitute the kind of representation which could be output from auditory preprocessing, or, if not, whether the auditory features output by the preprocessor could readily be translated into either phonemes or syllables. The debate within Psycholinguistics continues, and the evidence is mixed. On the one hand, there is by now a fairly substantial body of evidence that the syllable is a natural segmentation unit, at least for French (see Mehler, 1981, or Segui, 1984, for a review of this evidence). But syllabic segmentation effects which have been demonstrated in the recognition of French do not appear in the recognition of English (Cutler, Mehler, Norris & Segui, 1986; Norris & Cuder, 1988). For English, Pisoni (1981; see also Pisoni, Nusbaum, Luce & Slowiaczek, 1985) has argued that phonemes are the most useful segmentation units.

Other units have been proposed by speech engineers and psychologists in recent years; these include units both above die phonemic level (e.g. demisyllables: Fujimura & Lovins, 1978, 1982; diphones: Klatt, 1979) and below it, (e.g featural representations: McClelland & Elman, 1986; spectral templates: Klatt, 1979). It is generally the case mat models of auditory word recognition which have assumed a level of representation in terms of a linguistic unit such as the phonological feature (McClelland & Elman, 1986), the phoneme (Marslen-Wilson, 1980; McClelland & Elman, 1986) or the syllable (Mehler, 1981; Segui, 1984) have arisen from within cognitive psychology, and have not been directly concerned with questions of recogniser implementation. Non-linguistic units such as diphones (Klatt, 1979) or demisyllables (Fujimura & Lovins, 1978, 1982) have largely been proposed by researchers who are concerned more with implementation than with psychological modelling.

G. The Universality Issue

In the above discussion a simplifying assumption has been adopted, namely that the three levels of representation considered, auditory representations output by the preprocessor, input representations to the lexicon, and intermediate representations if any, will be the same for all speech perception operations. This is not necessarily the case. Precisely in the area covered above there exists considerable variation across languages. For example, there is variation in what may potentially constitute a lexical unit, whereby relatively uninflected languages such as Chinese contrast with highly inflected languages such as Turkish. Similarly, there is variation in the potential characteristics of lexical input representations. Here there is a major distinction in the domain of prosody, between languages which use prosody to distinguish between lexical units and languages which do not. The former group includes tone languages such as Chinese and Thai, and lexical stress languages such as English and Russian. The latter group (which is larger) includes fixed stress languages such as Polish or Hungarian, and all non-tone non-stress languages such as French. Finally, there is considerable variation across languages in the variety and characteristics of the linguistic units which are presented as viable candidates for prelexical representation. The number of vowels in a language can vary from as few as three to as many as twelve (English has more than twice as many vowels as Japanese, for example). Syllable structure can vary from languages which allow only or almost only consonant -+ -vowel syllables (Japanese is one of the latter, for instance) to languages like English, in which syllables may be as different in structure as *a* and *strange*, and in which stress patterns result in a wide discrepancy in acoustic-phonetic clarity between the realisation of stressed and unstressed syllables. Syllable boundaries, likewise, may be phonologically distinct (as they are in languages with regular syllable structure, for instance Japanese) or indistinct (as they are at the on set of many unstressed syllables in stress languages like English).

These sources of variation allow for the possibility that the very nature of the linguistic material to be processed may affect the way it is processed. Psycholinguistic models of word recognition have paid little attention to this possibility. Again, it is perhaps the bias of lexical modelling towards the visual domain which has obscured relevant cross-linguistic variation (though recently psycholinguists working in the visual domain have begun to examine the possibility that the nature of an orthographic code can affect the nature of the reading process - see Henderson, 1984).

There is a sense in which the interests of the cognitive psychologist here parallel, in a fortuitous but potentially productive way, the interests of the designer of a practical speech recogniser. The cognitive psychologist is concerned with the nature of the human recognition system, rather than the nature of the recognition system for any particular language. The striking characteristic of the human language acquisition system is that it acquires any natural human language with equal success; the mental capability of a newborn child, irrespective of its parentage, is not biased towards acquisition of one language rather than another. Thus if there prove to be language-specific variations in such aspects of speech recognition as the nature of prelexical representations, the cognitive psychologist is concerned to distinguish what is necessary to the recognition system from what is possible, i.e. to distinguish what is universal to the recognition process in all language users from what is specific to processing by users of a particular language. Universal features will be obligatory components of a model of human language processing; language-specific variations will comprise a repertoire of optional components from which the processor will select those components which best cope with the nature of the input.

In a similar way, the designer of a recogniser may employ knowledge of universal versus language-specific characteristics of the human recognition process to constrain the architecture of a system, by focussing on the design of those components which are universal to all human language processors.

Cross-linguistic study of auditory recognition within Psycholinguistics is in its infancy (Cutler, 1985). Very recendy, however, evidence has been found for a cross-linguistic difference in speech segmentation strategies, which may in turn imply a corresponding difference in the nature of prelexical or lexical input representations; Cutler, Mehler, Norris and Segui (1986) have produced evidence that the syllable is an effective segmentation unit for French but not for English. This suggests that psycholinguists may indeed need to develop a larger language-universal framework within which such results can be viewed as language-specific options. There is, however, substantial evidence that human listeners can make effective use of prelexical representations, of one kind or another.

H. Conclusion

The questions currently at issue in the study of human speech recognition concern the relationship between the output of the auditory preprocessor and the input to the lexicon. How can auditory features be extracted from the parallel auditory stream; how can such a representation in terms of auditory featuresbe segmented for presentation to the lexicon; how can it be transformed into a more abstract form corresponding to stored representations; does the transformation process necessarily imply an intermediate level of prelexical representation; and if so, in what order do segmentation and transformation occur?

Up till the present time these questions have not been the most central in Psycholinguistics. They have been comparatively neglected simply because of the separation of psycholinguistic terms of reference from those of auditory processing. Only the rapid growth of research on automatic speech recognition has encouraged psycholinguists to address these issues, because they must be resolved before the degree of relevance of human recognition evidence to the design of automatic recognisers can be determined.

However, the possibility of language-specificity at this level of processing is a dimension which should not be ignored. It is likely that psycholinguistic work will in the future become more cross-linguistic, i.e. will look at auditory word recognition and the segmentation and representational unit questions in the light of the ways in which languages differ. Such factors as presence versus absence of stress, relative occurrence of vowel reduction, frequency of prefixing versus suffixing, occurrence of stem-initial phoneme mutation, and phonetic functions of the prosodic dimensions of pitch, intensity and duration are all factors relevant to prelexical speech processing. It is at this level that the contrast between the psychoacoustic and the psycholinguistic approaches becomes particularly apparent. Psychoacousticians must be justified in assuming that the human auditory system is the same for everyone, and that the output of auditory preprocessing is the same kind of representation for all languages. Psycholinguists can no longer assume that the prelexical transformation process is the same for everyone, or that its output, i.e. the lexical input representation, is the same for all languages. Nonetheless, psycholinguists' new awareness of the transformation from auditory features as a central problem in speech recognition suggests that we may soon be seeing co-operative research projects addressing human speech recognition from the first auditory percept all the way to the lexicon. Such projects should, we suggest, also be of enormous value to engineers working on automatic speech recognition. In Pan 3 we suggest some techniques which might be exploited by this new research axis.

3. CONVERTING THE AUDITORY STREAM INTO A PHONETIC CODE

This pan of the paper outlines three current engineering approaches to the problem of converting the parallel data stream flowing from the auditory system into a sequence of discrete speech events. In each case, the acoustic input is subjected to a spectral analysis like that of the spectrogram and the resulting data stream is used as a substitute for auditory analysis. The frequency dimension is divided into channels and the number of channels varies from around 20 in vocoder style front-ends to 128 or 256 in the case of FFT-based front-ends. The temporal dimension is divided into time bins, or frames, which vary in duration from around 10 to 40 ms. The methods for generating the spectral representation vary considerably, but in each case, the data rate is relatively low and the temporal resolution is coarse in comparison with that of the auditory system. A detailed description of the techniques is presented in Bristow (1986); the current description is primarily concerned with how each approach tackles the problem of segmenting the parallel auditory stream into a discrete stream of phonological units, and to what extent each approach can capture cognitive psychological distinctions.

A. Feature Extraction

The traditional signal processing approach is based on the concept of feature extraction. Each frame of the spectrogram is searched for concentrations of energy, and adjacent frames are compared to establish the temporal and spectral extent of these auditory events. The events form patterns referred to as auditory featureswhich are often characteristic of the sound source. A subset of the auditory features that appear in the spectrogram represent speech events. For example, the pattern of formants that represent the [ae] in "past" and the burst of noise that represents the [s] in "past" (see Figure 1), are both examples of auditory features which are also speech events. In feature extraction models, the recognition system uses the features to establish the presence of phonemes, or other phonological units; then from this discrete stream of phonological units is generated a restricted list of word candidates with associated probabilities. Examples of different approaches to the feature extraction technique are provided by Assmann and Summerfield (1986), Duncan, Dalby and Jack (1986) and Lindsey, Johnson and Fourcin, (1986).

One of the main problems with the feature extraction approach is that it offers no particular solution to the segmentation problem. As we saw above, boundaries between units at all levels of analysis can be very unclear. There is no obvious cue either in the acoustic stream or in the auditory stream to signal where one lexical unit ends and the next begins; and the same is true of prelexical units. A portion of the signal which psycholinguists, and listeners, would unhesitatingly classify as containing two distinct phonemes, for instance, might offer no such clear contrast in terms of auditory features. As an example, a prevocalic stop consonant can appear more as a modification of the vowel that follows it than as a distinct auditory feature. Thus in the feature extraction approach the processes of extracting the features and of segmenting the continuous signal interact, and the approach therefore does not lend itself to a separation of levels of processing such as we have argued must be characteristic of human speech recognition.

B. Template Matching

In this technique, instead of each frame of the auditory stream being analysed separately, the frames are analysed in groups to see if the group contains a pattern that is characteristic of a speech event. It is a pattern recognition process in which the pattern in the group of frames is compared to each member of a set of canonical patterns, or templates. In fact the templates usually correspond to words, and so the template that provides the best match identifies the word candidate without the need of any intervening level of representation.

Template-matching approaches vary in sophistication from those which seek an exact match for untransformed stretches of speech to those which can cope to some extent with variability. The most successful technique at this time is Hidden Markov Modelling (HMM) and most current commercial devices use some form of it (Moore, 1986). It is a statistical pattern-recognition technique for modelling time-varying sequences and as such is particularly appropriate for speech. Each "template" is an HMM and each has to be learned. That is, the machine is trained on a range of forms that a word can take, and the HMM of that word is then a template that attempts to capture the variability of the word as well as its average form.

Template matching solves part of the segmentation problem inasmuch as the templates span whole sequences of what would be separate featuresin the previous technique. As a result, segmentation at the prelexical level does not arise, and the problem of segmentation is restricted to the level of the word-size template. The template has to be aligned with the part of the auditory stream to which it is being compared, and then it has to be stretched or compressed in time to fit the sample. The combined process takes a considerable amount of computation, and so, indirectly, segmentation remains an area where improvements are required (Cook and Russell, 1986).

The fact that a template is required for each word to be recognised means that there are far more primitive units in this system than there are in a feature-extracting system. And the fact that each template has to be compared with each input sample as it comes along means mat a recognition machine based on this technique requires considerable computer power if it is to operate in real time. Nevertheless, the technique provides impressive performance when compared to its predecessors.

C. Learning Machines

The final technique is connectionism, or neural networking. The technique arose in cognitive science as a development of learning-machine research. Recendy, it has been introduced into speech recognition as a means of converting the auditory stream into a phonetic stream (Bridle and Moore, 1984). At the same time it has captured the attention of psycholinguists as a useful framework for modelling human recognition performance (McClelland & Elman, 1986). In essence, a connectionist model is set up to learn the relationships between auditory patterns and phonetic codes. Many simple calculation units are set out in layers and each unit in one layer is connected to all of the units in the next layer by weighted links. Typically, units are connected to other units in the same layer only by mutually inhibitory links. In the case of speech recognition, the model usually has three layers of units: input units which characterise the auditory possibilities, output units which characterise the phonetic possibilities, and hidden units which connect the input and output units and make it possible for the model to learn complex relationships between the input and output states. The models are trained, as one would expect, by presenting the auditory patterns associated with words to the input units, the phonetic representations of the words to the output units, and adjusting the weights that connect the units to provide the "best fit" (Elman and Zipser, 1987; Landauer, Kamm and Singhal, 1987; Peeling and Bridle, 1986; Prager and Fallside, 1989).

The computation time taken to learn the relationship between a relatively modest set of auditory and phonetic events is currently astronomical: hours on a large mainframe computer and days on a workstation. However, once the network has learned the items, it can provide a phonetic transcription for an auditory pattern reasonably quickly. Part of the reason is that the network does not compare the input to all possible outputs sequentially. The memory in the network is contained in the set of weights derived in the learning session, and that one set of weights is used to convert all inputs to all outputs. The advantage of these machines, then, is that they effectively compare the input pattern to all of the stored representations simultaneously.

connectionist models have had similar problems to HMM models with respect to segmenting the auditory stream and scaling the stream temporally. In one recent model, Waibel et al. (1987) attempt to solve part of this problem by expanding the input-unit layer to include several copies of the current auditory input. It increases the architectural complexity and the computational load considerably but it does make the model more resistant to temporal variation. Very recently there have also been attempts to explore connectionist architectures which are specifically adapted to dealing with temporal information, for instance dynamic nets (e.g. Norris, 1988). These approaches will probably produce the next generation of connectionist recognition systems.

It remains to be seen whether this approach will lead to better performance than the HMM approach, connectionist modelling, does, however, illustrate how cognitive science is being extended into the realm of peripheral auditory processing. Importantly, it is also the first modelling framework to gain equal popularity with speech engineers and cognitive psychologists. Thus it offers, for the first time, a ready-made framework within which constraints derived from our knowledge of human recognition performance can be applied to the design of an ASR system.

CONCLUSION

In this paper we have described current work on the psychological modelling of auditory processing and word recognition. We have also briefly discussed available methods for connecting the auditory and speech systems, all of which now leads us to argue for a particular approach to the study of speech recognition, one which we believe offers the best chance currently available for new progress in the design of a general purpose automatic speech recogniser.

We have made two distinct claims. Firstly, we have argued that ASR research should make use of the resources offered by cognitive psychology. Although we do not yet understand human speech processing in sufficient detail to model the system both accurately and completely, we do understand a number of the constraints which apply to human processing, and in particular we know a great deal about the distinct levels of processing involved. The human speech recognition system demonstrates that real-time speaker-independent large-vocabulary recognition is possible. In the long term, therefore, the human system is both the standard which ASR seeks to emulate and, we would argue, the best model it can hope to adopt

Our second argument concerns the relationship between areas within cognitive psychology. Traditionally psychoacoustic studies of auditory processing and psycholinguistic studies of speech recognition have been independent and non-interacting disciplines. We believe that if cognitive psychology is to make a useful contribution to ASR research, cognitive psychologists first have to achieve an integrated model of human speech recognition which covers all aspects of the process from initial processing of the incoming waveform to successful location of stored representations of words. This means that psychoacousticians have to consider the nature of their model's output representation, and how such a representation might be constrained by the nature of subsequent processing; and it means that psycholinguists have to consider likewise the nature of their model's input representation, and how this can be translated into the discrete units required by the word recognition system

In the main body of the chapter we have argued that cognitive psychological modelling is relevant to ASR research, and that collaboration between psychoacousticians and psycholinguists is feasible. In Part 3 we suggest that at the present time one type of methodology presents the best opportunity for progress. connectionist modelling offers the prospect of uniting psychologists and engineers because it is a technique which is currently proving useful in both fields. It is also explicitly based, in a sense, on the human system in that the design of connectionist networks is intended to mimic the relationship between groups of neurons in the brain. We should make it clear, of course, that we do not consider this aspect of connectionist methodology to be central to its value; it is by no means necessary that a connectionist model is *ipso facto* a model of the human system. What we consider important in the present context is the computational power of connectionist systems, as well as the fact that they are adaptable both to cognitive modelling and to engineering design.

Current connectionist models of speech recognition, however, are implausible models of human processing. Consider the top part of Figure 8, which represents a typical current model. It has two stages: the first converts the incoming waveform into an auditory representation in terms of a spectrogram; the second is a giant undifferentiated connectionist model which attempts to associate spectral representations with words. As we have argued above, the spectrogram does not even approach the level of fine-grain analysis which the human auditory system applies to incoming waveforms. And as we have also argued, the conversion of auditory features to lexical representations in the human recognition system is not an undifferentiated process, but consists of a number of separable processing levels.

We propose, therefore, that the connectionist modelling required for the next generation of recognition machines should be more like the bottom half of Figure 8. Firstly, instead of relying on a poor-definition spectrogram, the system should simulate the human auditory system, mimicking first the processing which is performed by the cochlea, then the processing performed by the neural auditory system. Secondly, the conversion of auditory features to lexical representations should not be attempted in one stage; rather it should proceed in isolatable stages, involving intermediate levels of representation prior to lexical access.

This proposal does not, of course, constitute a complete and detailed model of the human system. For instance, the figure is explicitly neutral with respect to the nature

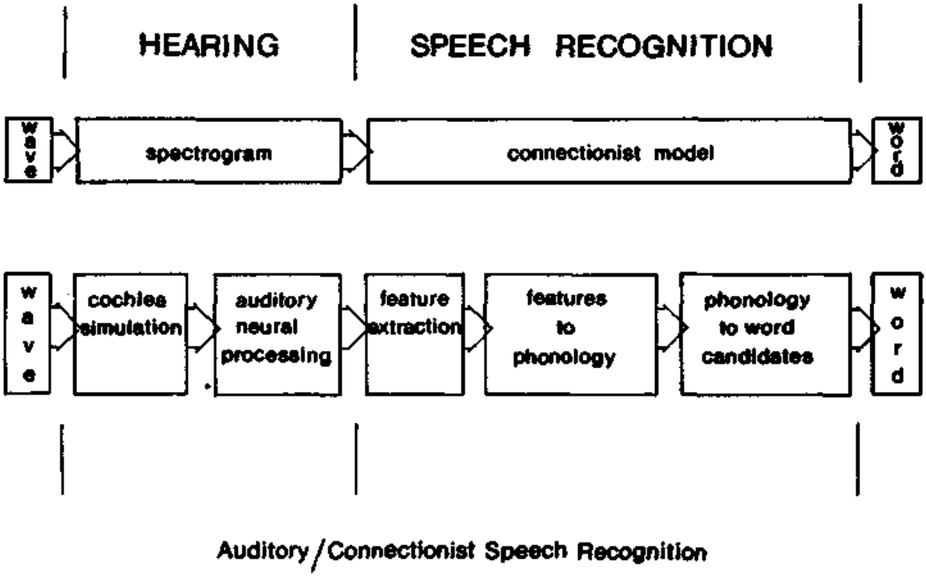

FIG. 3.8 A comparison of existing (upper row) and proposed (lower row) methods of word recognition using the auditory/connectionist approach. The spectrogram in the upper row is replaced by a full cochlea simulation and a pulse ribbon model of auditory neural processing in the lower row. The monolithic connectionist model in the upper row is replaced by a psychological, staged model in which features are extracted from the auditory image and converted into a sublexicai form of phonology before the phonology is assembled into word candidates.

of prelexical representations (phonemes, demisyllables and syllables are among the possibilities here). It is not the processing details that we are arguing for, it is the general structure of the model. We believe that this general structure is the right choice for the next generation of speech recognition models.

ACKNOWLEDGEMENTS

The implementation of the gammatone filterbank, and the software used to produce the figures in the paper, were both developed by John Holdsworth.

REFERENCES

Aitchison, J. (1987) *Words in the Mind: An Introduction to the Mental Lexicon.* Oxford: Blackwell.

Assmann, P. & Summerfield, Q. (1986) Modelling the perception of concurrent vowels. *Proceedings of the Institute of Acoustics: Speech and Hearing, Vol 1.* 8, Part 7,53-60.

Beet, S.W., Moore, R.K. and Tomlinson, MJ. (1986) Auditory modelling for automatic speech recognition. *Proceedings of the Institute of Acoustics: Speech and Hearing,* Vol. 8, Part 7,571-580.

Bek&y, G. von (1960) *Experiments in Hearing.* New York: McGraw Hill.

Bridle, J.S. & Moore, R.K. (1984) Boltzmann machines for speech pattern processing. *Proceedings of the Institute of Acoustics,* Vol. 6, Part 4,315-322.

Bristow, G. (Ed.) (1986) *Electronic Speech Recognition.* London: Collins.

Carlson, R., GranstrOm, B., and Klatt, D. (1980) Vowel perception: The relative perceptual salience of selected acoustic manipulations, *Speech Transmission Laboratory Quarterly Progress and Status Report* (TRITA-TLF-79-4), Stockholm, Sweden, 73-83.

Cook, *AM.* and Russell, MJ. (1986) Improved duration modelling in hidden Markeov models using series-parallel configuration of states. *Proceedings of the Institute of Acoustics: Speech and Hearing,* Vol 8, Part 7,299-307.

Cooke, MP. (1986) Towards an early symbolic representation of speech based on auditory modelling. *Proceedings of the Institute of Acoustics: Speech and Hearing, V*ol. 8, Part 7,563-570.

Cutler, A. (1985) Cross-language Psycholinguistics. *Linguistics,* 23,659-667.

Cutler, A., Mehler, J., Norris, D. & Segui, J. (1986) The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language,* 25,385-400.

de Boer, E. & de Jongh, H.R. (1978) On cochlear encoding: Potentialities and limitations of the reverse correlation technique. *Journal of the Acoustical Society of America,* 63,115-135.

de Boer.E. (1983) No sharpening? A challenge for cochlear mechanics. *Journal of the Acoustical Society of America,* 73, (2), 567-573.

Delgutte, B. (1980) Representation of speechlike sounds in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America,* 68,843.

Dolmazon, J.M. (1982): Representation of speech-like sounds in die peripheral auditory system in light of a model. In R. Carlson and B. GranstrOm, (Eds.) *The Representation of Speech in the Peripheral Auditory System.* Amsterdam: Elsevier; 151-163.

Duncan, G., Dalby, J. & Jack, M.A. (1986) Star-pak: A signal processing package for acoustic phonetic analysis of speech. *Proceedings of the Institute of Acoustics: Speech and Hearing,* Vol 8, Part 7, 77-84.

Elman, LL. & Zipser, D. (1987) Learning the hidden structure of speech. Institute for Cognitive Science, UCSD, California Report 8701.

Friederici, A.D. & Schoenle, P.W. (1980) Computational dissociation of two vocabulary types: Evidence from aphasia. *Neuropsychologia,* 18,11-20.

Fujimura, O. & Lovins, J.B. (1978) Syllables as concatenative phonetic units. In A. Bell & J.B. Hooper (Eds.) *Syllables and Segments.* Amsterdam: North-Holland.

Fujimura, O. & Lovins, J.B. (1982) *Syllables as concatenative phonetic units.* Indiana University Linguistics Club.

Gardner, R.B. and Uppal, M.K. (1986) A peripheral auditory model for speech processing. *Proceedings of the Institute of Acoustics: Speech and Hearing, V*ol. 8, Part 7, 555-562.

Ghitza, O. (1986): Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech & Language,* 2, 109-130.

Goldstein, J.L. (1973) An optimum processor theory for the central formation of the pitch of complex *tones. Journal of the Acoustical Society of America,* 54,1496-1516.

Goldstein, J.L. & Srulovicz, P. (1977) Auditory-nerve spike intervals as an adequate basis tor aural frequency measurement. In E J. Evans and JP. Wilson (Eds.) *Psychophysics and Physiology of Hearing.* Academic Press: New York.

Helmholtz, HX.F., von (187S, 1912) *On the Sensations of Tone.* English translation of 4th edition by A J. Ellis (Longmans, Green and Co., London, 1912).

Henderson, L. (Ed.) (1984) *Orthographies and Reading.* London: Erlbaum.

Houtgast, T. (1974) Lateral suppression in hearing. Thesis, Free University of Amsterdam, Academische Pers. BV, Amsterdam.

Hunt, M J. & Lefebvre, C. (1987): Speech recognition using an auditory model with pitch-synchronous analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP-87,* Dallas, 813-816.

Johannesma, P.I.M. (1972) The pre-response stimulus ensemble of neurons in the cochlear nucleus. *Proceedings Symposium on Hearing Theory,* pp.58-69. IPO, Eindhoven, The Netherlands.

Jusczyk.P.W. (1986) Areview of speech perception research. In Kit. Boff.L. Kaufman, &JP. Thomas (Eds.) *Handbook of Perception and Human Performance.* New York: Wiley.

Klatt,D.H.(1979) Speech perception: Amodel of acoustic-phonetic analysis and lexical access. Journal *of Phonetics,* 7,279-312.

Landauer, T.K., Kamm, C A. & Singhal, S. (1987) Teaching a minimally structured back-propagation network to recognise speech sounds. Bell Communications Research Report.

Lindsey, G., Johnson, M. & Fourcin, A. (1986) Diminution of high frequency energy as a cue to the voicelessness of following consonants. *Proceedings of the Institute of Acoustics: Speech and Hearing,* Vol 8, Part 7,25-30.

Lyon, R.F. (1984) Computational models of neural auditory processing. *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing* (March), paper 36.1.

Marslen-Wilson, W.D. (1980) Speech understanding as a psychological process. In J.D. Simon (Ed.) *Spoken Language Generation and Recognition.* Dordrecht: Reidel.

Mathes, R.C. & Miller, RX. (1947). Phase effects in monaural perception. *Journal of the Acoustical Society of America, 18,*780-797

McClelland, JX. & Elman, JX. (1986) The TRACE model of speech perception. *Cognitive Psychology,* 18,1-86.

Meddis, R. (1986) Simulation of mechanical to neural transduction in the auditory receptor. Journal of *the Acoustical Society of America, 79,*702-711.

Mehler, J. (1981) The role of syllables in speech processing. *Philosophical Transactions of the Royal Society,* B295,333-352.

Moore, B.C J. & Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America, 74,*750-753

Moore, R.K. (1986) Computational techniques. *Electronic Speech Recognition,* G. Bristow (Ed.), London: Collins, 130-157.

Norris, D.G. (1988) A dynamic net model of speech recognition. Paper presented to Workshop on Computational and Cognitive Approaches to Speech Processing. Sperlonga, Italy.

Norris, D.G. and Cutler, A. (1988) The relative accessibility of phonemes and syllables. *Perception & Psychophysics, 43,*541-550.

Patterson, R.D. (1973) The effects of relative phase and the number of components on residue pitch. *Journal of the Acoustical Society of America, 53,*1565-1572.

Patterson, R.D. (1976) Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America, 67,*229-245.

Patterson, R.D. (1986) Spiral detection of periodicity and die spiral form of musical scales. *Psychology of Music, 14,44-61.*

Patterson, R.D. (1987a) A pulse ribbon model of peripheral auditory processing. In William A. Yost and Charles, S. Watson, (Eds.) *Auditory Processing of Complex Sounds.* Hillsdale, NX: Erlbaum, 167-179.

Patterson, R.D. (1987b) A pulse ribbon model of monaural phase perception. *Journal of the Acoustical Society of America, 82,* (5), 1560-1586.

Patterson, R.D. and Moore, B.CX (1986) Auditory filters and excitation patterns as representations of frequency resolution. In B.C J. Moore (Ed.) *Frequency Selectivity in Hearing.* Academic: London, 123-177.

Patterson, R.D. Nimmo-Smith, I. Weber, DX. & Milroy, R. (1982) The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journal of the Acoustical Society of America, 72,*1788-1803.

Patterson, R.D. and Wightman, FX. (1976) Residue pitch as a function of component spacing. *Journal of the Acoustical Society of America, 59,*1450-1459.

Peeling, S. & Bridle, J. (1986) Experiments with a learning network for a simple phonetic task. *Proceedings of the Institute of Acoustics: Speech and Hearing,* Mil. 8, Part 7,315-322.

Perkell, J.S. & Klatt, D.H. (1986) *Invariants and Variability in Speech Processes.* Hillsdale, NX: Eribaum.

Pisoni, D.B. (1981) Phonetic representations and lexical access. Paper presented to the Acoustical Society of America, Ottawa, May; *{Journal of the Acoustical Society of America,* 69,532).

Pisoni, D.B. & Luce, P.A. (1986). Speech perception: Research, theory, and the principal issues. In E.C. Schwab & H.C. Nusbaum (Eds.) *Pattern Recognition by Humans and Machines.* Vol. 1. New York: Academic Press.

Pisoni, D.B., Nusbaum, H.C., Luce, P.A. & Siowiaczek, L.M. (1985) Speech perception, word recognition and the structure of the lexicon. *Speech Communication, 4,*75-95.

Prager, R.W. & Fallside, F. (1989). The modified Kanerva model for automatic speech recognition. *Computer Speech and Language, 3,*61-81.

Schofield, D. (1985) Visualisations of speech based on a model of the peripheral auditory system. NPL Report DITC 62/85.

Schwid, H.A., & Geisler, CD. (1982). "Multiple Reservoir Model of Neurotransmitter Release by a Cochlear Inner Hair Cell", *Journal of the Acoustical Society of America, 72,*1435-1440.

Seashore, R.H. & Eckerson, L.D. (1940) The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology, 31,*14-38.

Segui, J. (1984) The syllable: A basic perceptual unit in speech processing. In H. Bouma & D.G. Bouwhuis (Eds.) *Attention and Performance X.* Hillsdale, NX: Eribaum.

Seneff, S. (1984) Pitch and spectral estimation of speech based on auditory synchrony model, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* San Diego, (March), paper 36.2, vol. 3.

Shamma, S-A. (1986) Encoding the acoustic spectrum in the spatio-temporal responses of the auditory nerve. In B.CX Moore and R.D. Patterson (Eds.) *Auditory Frequency Selectivity.* Plenum: New York, 289-298.

Swinney, D.A. and Cutler, A. (1979) The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18,*523-534.

Taft, M. (1988) A morphological decomposition model of lexical representation. *Linguistics. 26,* 657-668

Terhardt, E. (1974) Pitch, consonance and harmony. *Journal of the Acoustical Society of America, 55,* 1061-1069.

Tranmuller, H. (1987) Phase vowels. In M.E.H. Schouten (Ed.). *The Psychophysics of Speech Perception.* Dordrecht: Martinus Nijhoff.

Van Lancker, D. (1975) Heterogeneity in language and speech: Neurolinguistic studies. *UCLA Working Papers in Phonetics,* 29.

Waibel, A., Hanazawa, X, Hinton, G., Shikano, K. & Lang, K. (1987) Phoneme recognition using time-delay neural networks, ATR Technical Report TR-1-0006.

Wakefield, G.H. (1987) Detection of envelope phase disparity. *Journal of the Acoustical Society of America, 81,* Suppl. 1, S34.

WegeL RX. and Lane, *CM.* (1924) The auditory masking of one sound by another and its probable relation to the dynamics of the inner ear. *Physiological Review, 23,*266-285.

Wightman, FX. (1973) The pattern transformation model of pitch. *Journal of the Acoustical Society of America, 54,*407-416.

Yost, W.A. and Watson, C.S. (1987) *Auditory Processing of Complex Sounds.* Hillsdale, NX: Eribaum.

Young, E.D., and Sachs, M.B. (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America, 66,* 1381-1403.