Monograph:
Kirby, H.R. (1988) How Can Speech Recognisers Help Applied Research in the Civil Engineering, Transport and Related Industries. Working Paper. Institute of Transport Studies, University of Leeds , Leeds, UK.

Working Paper 257

Published paper
Kirby H.R. (1988) *How Can Speech Recognisers Help Applied Research in the Civil Engineering, Transport and Related Industries.* Institute of Transport Studies, University of Leeds. Working Paper 257

Working Paper 257

January 1988

HOW CAN SPEECH RECOGNISERS HELP APPLIED RESEARCH IN THE CIVIL ENGINEERING, TRANSPORT AND RELATED INDUSTRIES

Report of a seminar at the University of Leeds on 5th November 1986 organised on behalf of the Environment Committee, Science and Engineering Reserach Council

H R Kirby

This seminar was sponsored by the Science and Engineering Research Council

HOW CAN SPEECH RECOGNISERS HELP APPLIED RESEARCH IN THE CIVIL ENGINEERING, TRANSPORT AND RELATED INDUSTRIES?

Notes of a Seminar
held at
The University of Leeds on 5th November 1986
organised on behalf of the
Environment Committee, Science and Engineering Research Council

Part A: SEMINAR OVERVIEW AND CONCLUSIONS

A.1 BACKGROUND

Speech recognition technology is rapidly advancing to the point where it can be usefully applied in a wide range of contexts. For applications within the SERC Environment Committee's area of interest - civil engineering, construction, building, transport, water resources there are a number of kinds of recording situation in which one needs to keep one's eyes on the situation being studied, or in which the recording conditions (eg moving around with instruments) are unfavourable. The limitations of conventional pen and paper recording for these situations are obvious, and the limitations of hand-held data capture devices are also becoming apparent. Speech is therefore an easier medium to use, and a tape recorder a convenient means of recording the observations. For well defined recording tasks, speech recognisers might be a helpful way of transcribing the record. This seminar was convened to enable those who are potentially interested in such an application of information technology to hear of the latest developments and assessments of the suitability of the technology.

A.2 OBJECTIVES

The meeting had three objectives.

(i) To review the state of art of speech recognition technology.

(ii) To review the scope for applying speech recognition technology in the building, construction, transport and water resources areas.

(iii) To consider whether there was a sufficiently strong level of interest within these areas to enable the Environment Committee to establish an SERC facility so that a range of academic groups could share the use of the equipment.

A.3 PROCEEDINGS

A.3.1 RThe role of the SERC

Geoffrey Eastwood (SERC) described why the meeting had been sponsored by SERC, and the nature of SERC's support for specialist facilities. His talk is reproduced at Part B.

A.3.2 State-of-art

John Holmes (formerly Director of the Joint Speech Research Unit and now and independent speech consultant) gave a state-of-art review of speech recognition technology, including illustrations of why speech recognition is difficult, in the first part of his talk. This is reproduced at Part C, Section 2.

The following manufacturer's / suppliers gave presentations about their equipment (in the absence of each other) followed by informal demonstrations:

- **Logica** (with LOGOS 1 and 2)
- **John Bradburn Associates** (with KURZWEIL 3000)
- **British Technology Group / PA Technology** (with ASR)
- **Marconi** (with MACROSPEAK)

Summaries of the characteristics and prices of these machines, and other salient points about the state of art, are given in Part C, Section 2. Those products demonstrated were chosen as illustrative of the state of the art rather than to be comprehensive: in particular, of American products, KURZWEIL was the only one demonstrated; other leading American manufacturers include VOTAN and AT&T.

A.3.3 Application areas

The second part of John Holmes's talk introduced the scope for applying speech recognition technology. This is reproduced in Part D, Section 1. **Peter Bonsall** and **Howard Kirby** (University of Leeds) reviewed some transport applications and summarised the results of recent work under an SERC grant (GR/C/69184) and a contract from the European Office of the US Army's Research, Standardisation nd Methods Branch. Details of the former were in a precirculated paper (Institute Working Paper 213). Part D, Section 2 summarises main points from their talk, together with main points from the ensuing discussion on areas of application in the civil engineering, transport and water resources areas. Overlaps with other disciplines' interests were noted: for example, recording of microscope observations of water samples, rock samples, medical samples would require essentially the same procedures for assessing the applicability of speech recognition technology.

A.4 PARTICIPANTS

There were 28 participants. These are listed in Part E, together with other interested parties to whom this report is being sent. Only two universities were represented, but three others are known to be interested in the possibility of speech recogniser applications. The low participation rate by academics was thought to be due to four factors: the change to a mid-term date; the lack of suitable 'umbrella' organisations through which to disseminate information (apart from in the transport area, which has the Universities Transport Study Group); and the fact that few academics had research projects in data-capture. However, other research organisations and end-users were clearly interested; a meeting set up more for these at the outset would have had a wide appeal.

A.5 CONCLUSIONS

From the discussions on the day the following conclusions became apparent.

(i) Functional capability. Subject to satisfactory controls to reduce problems due to background noise or other environmental problems, there now seemed sufficient evidence of speech recognisers' ability to perform limited recognition tasks to justify the development and detailed evaluation of demonstration projects in selected application areas, with a view to providing feed-back to the manufacturers on performance of existing equipment and desired functional specification for future equipment. The main factor limiting the performance of the machines was probably variations in speakers voice characteristics, due for example to stress or colds or variations in background noise levels.

(ii) Academics' interests. There was as yet insufficient evidence of commitment by academics throughout the transport, civil engineering, building and water resources area to justify establishing an SERC facility to serve just those academic groups in the Environment Committee's area of interest, though the case for evaluating the technology remained.

(iii) Researcher and end-users' interests. The commonalities of interest by some of the research and end-user organisations represented were sufficient to justify consideration of a different and wider-spead basis for enabling different users to have experimental access to a range of recognisers for evaluation purposes. It was recognised that such a facility would essentially serve a 'technology transfer' function; and it was proposed that SERC involve its Information Technology Committee in further consideration of the prospects for such a facility.

(iv) Suppliers' interests. It was recognised that the suppliers' interests would be best served with high volume sales; and that, insofar as this would bring the price down, this would also be in the interests of end-users. Whilst manufacturers might be concerned lest a shared facility might, in the short term, reduce the number of immediate purchasers, it was considered that, in the longer term, such a facility might increase sales as a result of successful demonstrations of their use. Some manufacturers/suppliers indicated a preparedness to consider making some special arrangement to help set up such a facility. The necessity of ensuring that the product and the user environment were well suited to each other was recognised. In some cases (eg use with masks on in sewers, use with remote microphone in cars), there would be special needs which would probably require a high degree of interaction between end-user and supplier to achieve the required performance.

(v) Extensions of scope. The possibility of convening a wider conference on the subject, with objectives more related to various application sectors generally than to the academic community, was canvassed, but this was thought to be too ambitious an aim at this stage.

Part B: THE CONCEPT OF AN SERC FACILITY

Geoffrey Eastwood (SERC Environment Committee) introduced the seminar with explaining why it had been convened, and in particular why the seminar is being sponsored by such an unusual body as the SERC Environment Committee which on the face of it has got nothing at all to do with speech recognition. Indeed, if it was asked to enourage speech recognition as such, it would refuse to do so quite firmly.

What the Environment Committee is concerned with is the support of civil engineering, building, transport operations research and, to some extent, marine technology. It already does that to a large extent in these four subjects, separately: civil engineers design beams, building people worry about heat losses from houses and so forth, transport people phase their traffic lights and the marine technology people work with oil rigs. However, the Environment Committee has been trying to develop an overall policy, and broaden itss scope somewhat, to see whether it could do something a little more enterprising than simply encourage those four subjects. It has therefore been trying to encourage a multi-disciplinary approach to its subjects, in the first place through co-operation between its various components as far as that is possible, but also by looking outside its own strict disciplines to see if it can learn anything from other people, and apply the techniques that other people have been developing.

One main strand of that policy is in the application of information technology. It has looked with envy at the vast sums of money which are being put into information technology research in the Alvey Programme, and has wondered whether it can capitalize on that, so it is proposing a programme of research on the applications of information technology in the construction and transport industries.

Secondly, it is proposing a generalised programme on the in-service renewal of infrastructure, which is concerned with replacing sewers and things like that which constantly seem to be collapsing under buses in towns like Manchester and Leeds. Whilst this is somewhat more inward looking these programmes will also look outwards as far as possible.

It is also concerned with providing important tools for people to use. This is really one of the bases of SERC. SERC was to some extent founded on the National Institute for Research in Nuclear Science and on the Royal Society's Rocket Programme, both of which are major facilities provided for universities, where the facilities were really too large for any individual university or research group to have for themselves. With that idea in mind, the Environment Committee has been providing major or special facilities for some of its users, in civil engineering, building, transport, and marine technology.

(a) Vibration table This was the first facility the committee set up, and is located at Bristol University primarily for work on earthquake engineering but also for vibrating anything that anyone would want to vibrate.

(b) Flood channel facility This is located at Hydraulics Research, Wallingford, for research on turbulent flow in shallow, possibly meandering, channels with the idea of being able to design flood protection works rather better.

Both of the above facilities are in the £0.5M bracket so they are quite expensive and significant.

(c) Anaerobic digestion facility This is being developed with the idea of treating strong industrial waste waters in the factory itself rather than just discharging the stuff into the sewers.

This is about £0.25M, and will be running on stream - in a very literal sense - early next year.

(d) Soft clay test site This is just about to be set up in Scotland, so that civil engineers can go and dig trenches in soft clay and stick piles into it and test it, so that they are sure about its geotechnical properties when they want to put a building on that kind of material.

These are all facilities which are very much within the scope of the community that they are addressed to. The final one is different:

(e) Artificial intelligence software facility This is looking outwards somewhat, because expert systems are one of the OK words these days in all branches of science and engineering. People want to use them. They are not very sure what they are, they have no experience of them and if someone puts in a grant application he tends to apply for two or three expert system shells because he has got no experience of any of them, and he therefore needs to work in a bit of expertise with the expert systems to start with before he even starts applying them to his own particular core subject.

In order to cut down these multiple applications, we are setting up this artificial intelligence software facility so that we will have a number of systems available which we can make available to potential users for an extended period of, say, two months. The investigator can then really try it out in his own laboratory conditions, really get used to it and know what it will do. If he goes along to an exhibition and has the thing demonstrated to him, it will be in terms of some rather simple-minded system which has been cooked up for the purpose, and he will come away saying "yes, yes isn't that wonderful, just what I want", and then when he buys one and gets it home he discovers that really it is all a bit more difficult than that. The system does not actually do exactly what he thought it did and it is probably a good deal more difficult to make it do anything at all than the salesman had led him to believe.

So this is a means of giving real in-depth experience of systems like this to potential users. Having developed all these ideas, the question then arose as to whether there was any scope for doing the same kind of thing with speech recognition systems; after all none of these core subjects have any direct involvement in speech recognition but there are researchers in these areas who are enthusiastic about the scope for applications of speech recognition to the problems which come up in, particularly in transport operations but also possibly in the other subjects. This seminar has therefore been organised in order to test the temperature of the water and to see whether there is any general interest in the application of speech recognition technology to Environment Committee type problems. If there is, then one might be able to put together the idea for a similar facility. This would make speech recognition equipment available to potential users so that they could evaluate it in some depth for a couple of months or so and really see whether it would meet their requirements or whether they were just wasting their time.

As to the chances of funding such a facility, we must recognise that SERCs finances are very volatile; sometimes we have a lot of money and sometimes we have no money at all. Whilst the idea of a speech recognition facility was formulated at a time when there was a £2m underspend, we are now overspent by £¦m, so there is no money at all. However, this need not discourage the initiative, because if there is a good idea which ought to be encouraged, we ought to formulate it, work it out in detail and then as times change and the finance fluctuates there might just be the opportunity of carrying it through. It will not of course be carried through easily, because the Environment Committee will have to be convinced that there is a really significant application for any facility which it approves; but it is up to those assembled today to investigate this possibility and to assess whether such a proposal could realistically be generated.

B.2 SEQUEL

Some positive moves to develop applications research using speech recognition technology may flow from this meeting, even though they may be different from the SERC facility originally conceived. Recipients of this report are invited to keep in touch with these, by contacting the author of this report.

Part C: STATE OF THE ART IN SPEECH RECOGNITION

C.1 WHY SPEECH RECOGNITION IS NOT EASY (J. Holmes)

The first part of this presentation is going to give you some background as to why speech recognition is a problem. This is a necessary preliminary, because after all we can all talk and we can all recognise speech; so people would obviously think that speech recognition is not difficult! The second part (reproduced in section D1) is going to consider the practical side of speech recognition, commenting on the sorts of things that existing types of equipment have been applied to and some of the problems and benefits that have come from those applications.

It is obvious that, for controlling machines or giving information to machines, speaking to them is a very attractive idea. After all, we can do it with humans; you speak to another human being and get information over very quickly, and if they are sufficiently subordinate you actually can control them by speech; so why should we not do the same for machines?

Now this appearance is a bit deceptive, and to illustrate the point I am going to explain something about the actual problems of human communication.

What a human does when communicating by speech is to get a concept from your brain to that of the listener. You have an idea that you want to tell somebody something. You formulate that concept into words, then somehow or other the properties of those words control how you move your vocal organs (illustrated in Figure 1); you can move the tongue, the vocal chords can vibrate, lips can move and so on; and what comes out is sound. This radiates from the mouth, goes down some communication path (which could in certain circumstanes have a telephone in it) and finally the sound impinges on the person's ear; and then there is a decoding of this sound.

Figure 1 here ......

Now one of the first things to emphasise is that it is not very easy to separate words from the concept, because when human beings think they tend to think in words. It is very dificult to have an idea that is not tied to the words of a language. Whilst many animals must be able to think without using words, the complexity of their thinking processes are probably rather more limited. Humans though need some internal code for handling things, and this tends to be the code of ones own language. So it is very difficult to separate the task of recognising the words from that of recognising the concept.

The real problem is the question of how the words are coded into the acoustic signal, because the acoustic signal of course is just a wave form of how sound pressure varies with time.

Figure 2 gives an idea of what that might be like. Yet we know what these words are, and there is obviously a relationship between the wave form and these words; yet if you have another person saying the same words, that wave form would look quite different. It would have certain things in common that can be revealed by fairly complicated analysis. There will be superficial familiarities in form yet substantial difference in detail, just from another enunciation even of the same word.

So is there a better way of looking at it? Well, complex sounds can be broken down into separate frequency components, just like light can be displayed as a spectrum. The analysis of sound, in terms of how much energy there is at various frequencies, is in fact a useful way of understanding it, and there is a method of displaying speech known as a spectrogram which is even more helpful. Figure 3 shows three words on the spectrogram, the words "shout out loud". The little bit that we saw the wave form in Figure 2 was roughly from point A to point B on that picture, so is on a very much more compressed time. The vertical axis on the display represents frequency, the range in this case being from 0 to 4 kh. (There is some energy at higher frequencies but that is not displayed.) The horizontal axis represents time, about a second in all - three words like "shout out loud" take about a second going across the bottom. The amount of power at any frequency at any time is indicated by the blackness of the display. For example, the "sh" at the beginning is fairly weak, so is not very black, and the "sh" contains nearly all high frequency energy, hence most of the frame for "sh" is at the top, with nothing down at the bottom. Then, as the vowel starts, you get energy spread all over the frequency range. You will notice this trace has got vertical stripes. The reason for those vertical stripes is that the sounds like vowels are produced by the so-called vocal chords which are folds of tissue at the top of the trachea in the larynx that vibrate and let puffs of air through; and every puff of air through excites the resonant system of the mouth and throat and causes the column of air there to vibrate at certain characteristic resonant frequencies. Strictly speaking, the effect is due not so much to when that puff of air is coming, but when the puff of air is shut off. This happens when the vocal chord snaps sharply together, its shock excites the column of air, causing it to vibrate. In this way a lot of energy suddently builds up, and then dies away; and then in the next cycle of the vibration that happens all over again.

The other important point about traces in the spectrogram is the resonances. These are controlled in frequency by the actual dimensions of the vocal tract, the position of the tongue and so on. These resonances do of course mean that one is going to get areas of frequency where the frequency components in this shock-excited vibration are emphasised, so one gets resonances at these four frequencies; and as you move the tongue around, so they move around.

Figures 2 and 3 here

The purpose of displaying this particular spectrogram is to illustrate that certain things about these sounds are linguistically similar. For example, in "shout out loud", you have got "out" in both of the first two words; and you will notice that in broad terms those patterns are somewhat similar. Thus, the "ou" part, which is roughly from C to D in Figure 3 in each case, you can see this second resonance is moving down, not quite at an even rate, whilst the first resonance moves up and down again a little bit. The times for the two "ou" sounds thus have slight differences, but are broadly fairly similar. Now however look at the trace for the two "t" sounds. The "t" of the "out" is shown by the trace at C, which is preceded by a little bit of silence, denoted by the pale patch. Now look at the second "t"; you might say, well, the first one is "out", and the second one is "out", so they should both show the same characteristics. But in the second "out", there is something completely missing from the trace compared with the first "out". That is because it is followed by an "l", so: "out loud"; you do not actually make the same sort of noise "t" as you do when the "t" is followed by another vowel. You can envisage this for yourself by thinking of the sound "out loud". In one case you have got a clear "t" and in the other place you have got more of a "kl", and it is quite a different sound. Sure enough, the trace of this noise shows no high frequencies visible at all.

The next thing to notice in the diagram is the difference between the traces for "oud" and "out". These two sounds, the "ts" and "ds" are both made with the tongue in roughly the same place in the mouth, pressing up behind the teeth ridge, so you would think that the acoustic properties would be similar. Well, they are in the sense that the frequencies of these resonances at the end of the sound are roughly the same at E as they are at F; but look at the difference in length. To a large extent that difference is because this "ou" is followed by a "d". A different type of consonant actually alters the length of the preceding vowel. It is also altered by the fact that it is the last word in the sentence, which tends to make it longer; these two factors both tend to make the second "ou" sound longer than the first "ou" sound, so you can see the sorts of differences that occur as well as the similarities.

At least in this spectogram you can see three quite nice clear boundaries, so you might think it is easy to perhaps separate out the three words. I wish it were always so. Even here, there is the question of not knowing which word that consonant belongs to, for example, is it "shoutout", which obviousy does not make sense, though apart from that problem at least you can see some separation. But what about a sentence like "How are you?".

Now we think of that sentence as having an "h" at the beginning and an "ow" sound and an "r" sound and a "y" and an "ou"; but when you actually think how you say it, imagine saying it very slowly, "h-o-w a-r-e y-o-u". There are no boundaries at all, just continuous change, and in fact if you look at a spectrogram of that you will see what the problem is (Figure 4). There is just no way in which you can see some boundary there. For example, where is the boundary between the end of the "ow" sound and the beginning of the "are"? We just do not know. It is not precisely defined.

So these are the sorts of problems that we have. Now what we want to do for automatic speech recognition is to somehow recognize these individual sound components, which the phoneticians call the phonemes, the linguistic individual units. The problem is, how can we recognise the "aich" phoneme, say, when there is no precise way of identifying where it is, becaue phonemes clearly overlap and influence each other. We need to find some way of recognising the phonemes and so recognising the words.

Now it is not generally possible to recognise phonemes independently of identifying what the words are; - even human listeners cannot reliably identify what the phonemes are. If you get somebody making up some nonsense word that does not exist in your language and say it quickly you will very often not know what phonemes it has in it. There would be a certain amount of uncertainty about what it should be, and so it is really very difficult to recognise the phoneme without recognising the word.

So you might well say in that case how do we manage to recognise speech? Incidentally, this does not sound much different from " ...." which of course is another thickness of word with very similar phonemes and this is the sort of problem we have.

What we have to do is to make the interpretation of the sound make sense. We use our knowledge of the constraints and redundancies of the language to resolve a lot of these ambiguities. Figure 4 illustrates some of the linguistic effects that we can use.

<u>Figure 4</u> <u>Similar sounds, distinguishable by context</u>

Q. What's the weather like?	Q. What type of flour is this?
A. Grey day	A. Grade A

Consider these two different questions which might have these two answers. Unless the speaker of the answers was trying to distinguish between "grey day" and "grade A" because the speaker realised there was ambiguity, these two answers would probably be not reliably acoustically different at all. Yet they have actually got different words in. They have the same phonemes but different words. Yet in the unlikely event of somebody really asking one question or the other and getting the same answer, it would not even occur to the listener that there was any ambiguity because the wrong answer just would be so ridiculous as an answer to that question that it would not even be considered. Obviously, that was a somewhat contrived example, but that type of context-dependent interpretation is a really genuine effect.

Sometimes of course the context can not help, and in that case very small differences could be crucial. Another rather contrived example is shown in the two sentences in Figure 5.

Figure 5 Similar sounds, indistinguishable by context

Q. What would you like?

A. Some meals A. Some eels

These have got the same sequence of phonemes in a sense, except that in the first there are actually two "m" phonemes joined together, as "some" ends with an "m" and "meal" starts with an "m". Now acoustically the effect of that is to produce a longer "m", and bearing in mind there is no guarantee that speech sould will always be the same length on other occasions, you can certainly get some situations whre you could get somebody say that sentence and it would not be clear whether they meant "some meals" of "some eels". Obviously you can deliberately distinguish between them but there is a continuum of variation of length of the "m" where people will make the judgement one way or the other. Again, a contrived example but it illustrates the sorts of problems.

A common situation in which sounds are confused is when you try to spell out something over the telephone and you are going through the words one by one. You might, for example, say "T-H-E" and the listener writes down "T-A-G". The phonemes in this case are very slightly different. For "H-E" and "A-G" there is a slight phonetic difference between the phoneme that occurs round about the middle, but it is a minimal difference that is not reliably perceived on its own. When the phoneme is in actual words you do not have any problem at all. Anyone can converse over quite bad telephone lines and get the message over alright; it is only when somebody wants to give you an address or something like that and they have to spell it out that there is a risk of confusion, unless they give you key words for some of the letters.

Well that was the pessimistic side of the speech recognition problem. Now to go on to the good news as far as engineering-related applications are concerned. We consider first, what can be done by simplifying the problem

There are various things one can use for this. Sound patterns, for example spectrograms, or time/frequency intensity plots. Repeats of the same word are going to be more similar to each other than they are for different ones. That is assuming the words are chosen sensibly. There is some way forward therefore, by trying to match the sound pattern of complete words, and accepting that one has some limited vocabulary of words that are sensibly chosen so that they tend to be different. The type of machine that you would have is something like that in Figure 6, where you have some store of patterns, speech input coming in from the right, a pattern matcher that compares the pattern of speech coming in with all the patterns in the store, and the one that is best yet is output as the text sequence of words. So that is what you can do by pattern matching of whole words.

First of all let us assume that the words are spoken in isolation, so at least you get over the problem of working out where the word boundaries are by saying that they are when silence stops or starts. If you can do that reliably - not that it is always easy - it does not seem too bad a method. Mind you, there are clearly cases where this will not work. I mean take an example of a word like "can", and another word like "cam", with an "m" instead of an "n" at the end. Now as normally spoken they would be about the same speed, the actual acoustic difference between those two consonants, the "n" and the "m", is really quite small, so it can be quite difficult to reliably distinguish between them. But even so there is a fairly consistent time and acoustic difference so maybe it would not be too difficult to distinguish them; except that when you have given examples into your store you have said some of the words quickly and some of them slowly. Suppose you have said the "can" quickly and the "cam" slowly, and then the one you had to recognise was "can" but said slowly. If you are trying to match this incoming slow "can" with a quick "can" and a slow "cam", which one is it going to match best? It will probably actually match the slow word better because the difference in length might cause more difference in the pattern than the quite small difference in the acoustic properties of the consonant. Clearly then you would avoid choosing words that had that type of difference, you would choose words that were acoustically very different and when you had done that it is not so difficult to distinguish them.

However, isolated words are inconvenient. People do not speak words in isolation normally. Simple commands, yes; but if you wanted to give someone a telephone number you would not say "3-6-3-2-8", you would say "36328", in which case there is no gap, and no way of defining where the word boundaries are. As in the "how are you" example, there is no clearly defined gap, so how do we do it?

Well, there are obviously two problems with connected words using the pattern matching type of approach. One is the fact that the patterns themselves that are coming in are modified by their neighbours, as in the "how are you" example. In that example there was a continual change between the end of "how" and the Figure 6 here beginning of "are"; and obviously that word "how", if followed by a different word, would actually have its pattern altered.

Another problem is that you do not know where the boundaries are, as explained above.

Now strangely enough the second problem is actually solved. The more advanced speech recognisers that are on the market today use a rather clever computational algorithm in the pattern matching process, that considers all possible time places where there might be a boundary between words. At first sight you might throw up your hands with horror and say that such an approach would be very very demanding in terms of the amount of computation involved. But a rather clever algorithm has been designed that can do that; it considers all possible interword boundaries for matching the patterns without an astronomical computational increase. Whilst pattern matching for connected words is more complicated than for an isolated word recogniser, it is still quite practicable to do it, and not too difficult. So the problem about not knowing where the boundary is is solved.

The other problem about words influencing each other is, of course, a slightly different matter. But if the words are not too short, for example with words like "intelligibility", which are several syllables, it is only really the end syllables that are going to be influenced by their neighbours. It depends on what the beginning and end phonemes of the words are as to how much they are influenced. Some phonemes are influenced more than others. Vowels tend to be influenced a lot, some consonants tend to be influenced rather less. This means that by comparison with some standard pattern which might have been spoken in isolation, the patterns of the connected words will not match terribly well at the end of the word, though they will match reasonably well in the middle. Thus there is still a good chance of selecting the correct word from the vocabulary if you have chosen your vocabulary well enough. Most of the time this is what people actually do so connected word recognisers at the moment can actually work reasonably well.

We should now consider what the current capabilities are. There are now on the market several machines which can match the sound pattern for a limited vocabulary of stored words spoken by a human. This human, whose vocabulary is stored usually has to be the current user on the grounds that patterns being matched are more likely to be similar if the same person has spoken the word. Obviously, if somebody different speaks it, then there is more chance the patterns will be different. If the other person actually speaks with a completely different accent, of course, it is even worse; for example, if a northener says "grass" and a southener says "grass", the sound pattern is quite different, and you would not expect the two patterns to match terribly well. However, within the same accent area, the patterns are more likely to be similar, and if you use the same speaker that is a good thing to do if you can.

Speech recognisers that accept only isolated words tend to be a bit cheaper because they have not got to have the slightly more complicated boundary detection algorithm in the machine. More advanced systems still can actually accept continuous speech. Now the difference between continuous speech and connected words is this. Connected words means that if you want to say a telephone number you do not actually have to say each digit separately but it is still assumed that you recognise the utterance as a whole thing. In other words, if my telephone number is "36328", you recognise the whole utterance by starting to speak it and stopping speaking, and you recognise the boundaries of the utterance.

Now in continuous speech recognisers it is assumed that the person never actually needs to stop speaking before the recogniser makes a decision about what the beginning of the utterance is. In other words, as long as the person does not actually run out of breath, it is possible to design these connected word type of algorithm so that when you have got a little way into the script it has decided what the early words are. Obviously one has got to get some little way in, because if you start off with a word in an utterance like "six", you would not know whether the word was six or sixteen; so of course until you have got onto the second syllable and found it was not "t" you would not know that the first word was not sixteen. But say you did hear "teen" as the second syllable' that does not mean the word was "sixteen", because you might have been saying "six teenagers" ... So you can see the problems, the machine always has to be a little bit behind in its decisions. It cannot decide until you have spoken a word or two ahead of where it is making its decision, but once you have got a word or two ahead it can actually make a decision on the earlier words. It can do this continuously while you are carrying on talking. So that is the distinction between continuous and connected word recognisers, and there are machines available now which do just that, and do it quite effectively using this simple pattern matching algorithm.

The main issue left to discuss about current capabilities is the question of prices. Obviously, prices of electronic equipment depend very much on volumes of production. For example, pocket calculators are pretty complicated but they cost only three or four pounds becaue they are made by the million. If speech recognisers are only sold in fairly small quantities then there is not the amount of electronic development for real mass production, not only are they not so highly developed, but the manufacturers naturally want to recoup as much of the costs of research and development as they can, so they price them way above the manufacturing cost; they have to, because they have got to pay for the development.

Current prices range roughly between £1,000 and £10,000, mostly nearer the top end for machines of this sort. The isolated word recognisers are a bit cheaper, in general, and the more elaborate continuous speech ones are a little more expensive, but in that general range. The manufacturers information provided in Part 2 will no doubt give a much better idea of prices. There is no reason in principle if the market ever took off why the prices should not come down dramatically, simply because of the special development of large-scale integrated circuits. The sales volumes have got to be very high for doing that however.

Now as far as future performance is concerned, of course that is the subject of current research, and it is difficult to predict the outcome. But the sort of things that I think are going to happen are something much more like this slgithly different block diagram, shown in Figure 7. Here, instead of just a store of patterns, you have got some real elaborate linguistic knowledge, phonetic rules, as to the way sound patterns influence each other, the way the phonetic content of one word causes the next word to be said differently and so on. The acoustic analysis has got to be probably somewhat better than the simple spectrogram type of analysis that I showed earlier, because it probably has got to give the right sort of weight to those features of the acoustic signal that the human being does. The things that are phonetically important have got to be very important in a pattern interpreting process. There is a lot of research going on in this area, with many different approaches. It is completely outside the scope of this seminar, except to say that some improvements will have to come eventually, but they will take some little time.

Figure 7 here

C.2 Product descriptions

Part D: APPLICATIONS

D.1 Part 2 of the Keynote Address by J. Holmes

What sort of things can you use speech recognisers for? The advantages of using speech for controlling machines is if the person who is normally controlling the machine has got their hands or their eyes busy doing something else. A good example is parcel sorting. You are picking up a parcel, looking at the address, putting it on the conveyor belt and telling the machine where to route it. Obviously, if you have got to be simultaneously pressing buttons and looking at a keyboard you cannot be concentrating on picking up the next parcel. There is quite a big advantage there. Airline baggage sorting is exactly the same type of application, and United Airlines in the United States have operational speech recognition systems. I heard a talk by one of their representatives at a conference a few months ago where he said it would really be disaster if suddenly somebody took their speech recognisers away from them. It would completely alter the way they do their baggage sorting.

Another example might be if you are in a hostile environment. For example, if you are trying to control the machine out in the pouring rain and it has got a fairly subtle keyboarding method of operation, it is not very easy, whereas you can probably protect a microphone much easer than you can protect the keyboard or push-button control system. So there can be advantages in such a situation, or in situations where the person has got to wear gloves and cannot do the sort of subtle digital manipulation required for push-button or keyboard, but can still speak quite satisfactorily.

So those are examples of voice controlled machines. Now the need for the machine to be dedicated to the user, in other words to have the user's pattern for the machine, is not really a problem. You can train the machine by putting those patterns in fairly easily, because, even though you may have two or three operators using the same machine, you are not likely to have thousands and have new ones turning up at short notice. You will know in general what you are doing and who is doing it. So that is quite easy.

Data Entry Another application, quite interesting for engineering related problems, is data entry. Now one thing that you would not, I think, use speech for is in the average data preparation room where you have got loads of key punch operators just preparing cards or punching data straight into a machine. This is because if they have got the data before them on paper the trained operator can probably key it in quicker than they can read it, so I do not think a speech recogniser would be appropriate for that. What it does do is to stop the speech ever having to be written on the paper for the key punch operators to read. You actually collect the original data as speech, either going straight into the machine if that happens to be the operational environment you are in, or via a portable cassette recorder. If you have to be mobile like gas and electricity meter readers, you could have portable cassette recorders and just read the meter and speak the meter reading into their cassette recorder. Similarly you could do a traffic census with a portable tape recorder.

You then take it back, play it through the speech recogniser which then decodes it and puts it into the computer. A points of sale in retailing, the keyboard present could be used for putting more complicated information in, but you then have to learn the numeric codes for this, which is difficult. An easier alternative is to say "one pair of jeans, size so-and-so"; this is much easier than having to put in some stock control code. This form of data entry would make stock control systems much easier.

Production Line Inspection One application that occurred fairly early on in the history of the speech recognition industry was that of checking tolerances on television tubes on the production line. The person had to pick up the television tube and put it up against some measuring calipers and then had to say what the dimensions were. Now what that person could not do was to simultaneously key in those dimensions, because they were holding the measuring equipment or the tube.

Another example where speech recognition has been used is inspecting the trim on cars on the production line. Here the inspector is delving into the car, looking here and there, and very clearly cannot use a keyboard, but can use a microphone to report there is "a tear in the trim of the roof".

There is of course the question as to whether speech recognisers are accurate enough. Of course, the accuracy is not a hundred per cent. Speech recognisers do sometimes go wrong. Or it is not the speech recogniser that goes wrong, it is the speaker who does not speak clearly enough! The question is, what is the competition? Traditional methods involving the data prep room do make efforts to be accurate. The efforts may not be made by the actual keying operator, they may be made by a supervisor in reading bad writing on the form they are reading, they may be made by the person who is writing the number down putting it down wrong, and realising this between making other observations and so making a rule to correct his previous mistake. I think that quite alot of studies have been done that suggest that the accuracy of doing things by traditional methods is actually lower than is fairly easily achievable by a fairly good speech recogniser. So they can be really useful.

Access to Computer Databases Most people that are wanting access to some database have a VDU terminal and they key in the appropriate instructions to get the data displayed on the screen. Now that is a useful way of doing it, and the main advantages of speech are not for that. However, if you have a menu offered to you, asking which of five options you want, it may be quicker to say "one" than to find the key "1" and press it. Certainly, if there are questions, especially simple "yes/no" questions or two-way choice questions, for example as to whether you want profit or current bank balance, it would be very nice for us to have the machine speak and ask that question, and for the person to just reply with one or two words saying which of the two replies it is. That is the kind of thing that speech recognition can do very easily and it is actually quicker than displaying questions on the menu and indicating "press number one if it is option one and number two if it is option two". So speech recognition is quite good for that application.

But the main application for accessing a computer data base is if you have got telephone access to the database and you do not have special equipment. Then you have got to use speech, there is no other way around it if you have no special equipment. If you can put your request into the computer by speech in some way in some coded form, using a limited vocabulary, and get the answer back in the form of speech with voice output that is quite useful. But there is a problem. Usually you would want the recognisers capability to be speaker independent, if this is going to be useful, particularly via the telephone, you are going to have a wide variety of users whose voices you do not know in advance.

Now the effect of user independence on performance is of course that, in general, because the patterns are likely to be more different between different speakers than between different words uttered by the same speaker, you are not likely to have such a large vocabulary of words reliably recognised. You have therefore got to choose words that are more different from each other, to reduce the danger of speech from another speaker being similar enough to the wrong word to be mistaken for it. So much smaller vocabularies are required for speaker independence at the moment, until further research has solved that problem.

Interviews The next class of applications seems a little bit more far-fetched perhaps, involving interviewing, questionnaires and so on. Some people have done experiments using computers to ask people questions, with the keyboard used for the answer; but some firms have begun using speech. For medical diagnosis interviews, the doctor normally asks a lot of very probing questions about symptoms and of course sometimes if the particular medical condition is something of an embarrassing nature the person is a little reluctant to reply, even to the doctor. It has been shown that people do not mind replying to the machine as much, so if you can get the machine answering all these questions then maybe you can actually get the person to do so too - and what is more you save the time of the very sympathetic and specially chosen person who is capable of doing these interviews. That is a realistic application, and what is more of course you have got the answers already in the machine for putting into your expert system to help the diagnosis and determine what needs to be done about it. Of course, for this application, speaker independence is really essential, so you have got to design your dialogue very carefully indeed to enable only a limited vocabulary. For example, if you say to somebody, "do you have headaches in the morning", obviously the answer should be either yes or no. Clearly either you do or you do not, maybe "sometimes" might be a permitted answer. But what is most likely to happen with such a question is that the person will say "no, I get them at night", and as the machine would not be expecting that, you have really got to be very careful with the dialogue design to make the system work. But if you are careful enough - for example by giving the person sufficient instructions to start with to make them try and help, and by building in routines that will realise when an unexpected answer has been given, and will prompt the user to repeat his reply, saying for example "I'm sorry, I didn't quite understand that, could you..". You can do such things if you do it very carefully, but a lot of very carefuyl planning is needed.

Those are of course just a few examples. The examples are not exhaustive by any means, and I am sure that those present could envisage how one could expand on these in your own areas.

The next issue to consider is the problems of specifying the performance. Vendors of speech recognition equipment will very often quote percentage accuracy. This typically states 99% at such and such a vocabulary size, such a statement, with all deference to those vendors who are represented here, is absolutely meaningless, if stated with no other information because, whilst vocabulary size is of course important, so too is the quality of the speech. The sorts of things that are going to vary the accuracy are very much the manner of speaking and the actual selection of speakers. Whether there are connected or isolated words is going to make quite a big difference in the accuracy. Even if you have a connected word recognition algorithm, that is, one that will cope with connected words, it will generally give a much higher accuracy if used on isolated words. That is because, if you think of the "Grade A", "grey day" distinction referred to earlier, there is no danger whatever of confusing those two interpretations if the words are isolated; it is only when they are joined together that you do not know where the word boundary is. It is not that there is anything wrong with the algorithm, it is just that the actual words are spoken as a valid sequence of words with distinct word boundary positions.

Another issue is background noise. If you have got a really good, carefully designed acoustic environment, the speech is going to be much more reliably represented and the signal is going to be much more reliably representative of the phonetic content that you would find if there is a lot of clatter going on in the background.

Care in training is absolutely crucial. If you happen to have put your store of patterns in without enough care, no amount of care in speaking is going to solve your problem. I have actually seen people try and demonstrate speech recognisers where they trained it rather sloppily, just providing examples of the word quite quickly without being very careful. Repeated attempts to get a word recognised correctly may result in even worse matches, because the word is being said in a manner that makes it even more different from the training pattern; so you have got to be very careful.

So, important points. First of all, if you are going to do tests, use a standard speech database. You specify that it gives a certain performance on a very particular recording of speech that is available to the industry at large. If you are going to do that, it is most important that the developer should have not used that database for testing while he was developing the machine, because if he had done so and found certain errors on certain words, he might have tuned up the machine to solve those particular problems; but of course with another database it could perform completely differently. So you can make a machine that actually gets no errors at all on a standard database, but on some other standard database of similar quality with similar difficulties it might go wrong loads of times. You really have to be very cautious about that, and it really means you have got to be very cautious in assessing the machine. After all, if somebody says you get such and such an accuracy on this database, what I would want to know, and what those of you who are the central purchases, should want to know, is did the manufacturer do any development of the machine after he had tried it on that database. Because if they have, you must not trust that result, you need to test it on something else.

Now to the question of the design of a system. The speech recognition machine is commonly referred to as a black box, because from a device you put speech into, you get some word decisions. That is not the application. The application is the speech recognition machine in with everything else, and you have got to design the system to suit the application. If you design your system so that you are actually demanding from the recognition machine something that is achievable it can seem a marvellous boon and be extremely useful. Yet you might have a machine that is inherently much more capable, but you put it in a system that it is not really designed for. If you do not take everything into account, particularly human factors, the way people are going to speak and so on; and do not design the system in such a way as to make people speak in the right way, then it will actually be a complete failure. And the danger is of course, if you get into that situation then it gets the technology a bad name before its been tested properly. The danger would then be that fund-giving bodies like SERC would then say "oh no, we don't want that technology, it only causes trouble"; but that might arise simply by people not testing it properly, by not attending to the other factors, and then blaming the speech recogniser rather than the other factors.

What of the future? Speech recognisers are of very great benefit already for some. For most of the applications where there are enormous benefits, the cost is not critical. For example, for things like production line inspections, the saving in money by using speech recognisers is so great that even if they cost £10,000 it does not really matter.

For large-scale applications, for example data entry to office personal computers, the capabilities at present are not generally suitable. You can design special applications where it is worth it, but the cost-benefit analysis at the moment at present-day prices does not really add up, and this means of course that speech recognisers do not get used, you do not get the mass production, therefore you do not bring the price down. It is a vicious circle. Once the price comes down or the performance improves sufficiently to make the applications a bit wider, the sales volume will increase and of course the prices will come down. So I think it is bound to come. There will be, I really believe, an explosive growth; though there would be some delay in the purely financial aspects, the main delays will I think be due to the lack of research knowledge in making the performance good enough for some of the more general applications; that is, for the moment, the requirements for tailoring the speech to suit the machine is going to limit the range of applications. There are still loads of important applications but there is going to be a limit. When it becomes much more easy for people to speak much less carefully and less precisely to the machine; and when you can for example speak your letters or talk straight into the machine just like an office dictating machine, and get accurate enough performance out of it, then of course there will be the demand, and then everybody will benefit - except perhaps some of the present human operators who may be replaced!

Part E: PARTICIPANTS AND OTHERS INTERESTED

E.1 Participants

Mr Addy	Highways & Engineering Dept Leeds City Council Selectapost 5 Merrion House 110 Merrion Way LEEDS LS2 8JR Leeds (0532) 463000
Bob Baker	Transport & Road Research Laboratory Urban Networks Division Old Wokingham Road CROWTHORNE RG11 6AU Crowthorne 773131
Ian Blair	Highways & Engineering Dept Leeds City Council Selectapost 5 Merrion House 110 Merrion Way LEEDS LS2 8JR Leeds (0532) 463000
David Bradburn	John Bradburn (Computing Systems) St James Mill Road NORTHAMPTON NN5 5JW (0604) 55142
Geoffrey Eastwood	SERC Polaris House North Star Avenue SWINDON SN2 1ET 0793-26222 ext 2154
David Guyatt	Speech Systems Division Marconi Secure Radio Systems Ltd Browns Lane The Airport PORTSMOUTH PO3 5PH (0705) 664966

P Hawkes	BTG 101 Newington Causeway LONDON SE1 6BU 01-403-6666
Ray Heywood	Highways, Engineering & Technical Services Unit (HETS) Dudley House Albion Street LEEDS LS2 8JX Leeds (0532) 463000
John Holmes	Speech Technology Consultant 19 Maylands Drive UXBRIDGE Middlesex UB8 7BH (0895) 36328
Roy Jones	Marketing Analyst British Rail Eastern Region (Provincial) Room M73 York HQ YORK YO1 1HT York 53022 ext 2798
E S Judge	Water Research Centre (Processes) Elder Way STEVENAGE Herts SG1 1TH (0438) 312444
Ian P Leslie	Logica SDS Ltd Cobham Park Downside Road COBHAM Surrey KT11 3LX 01-637-9111 ext 2174

Christine Marsden

SERC
Polaris House
North Star Avenue
SWINDON SN2 1ET

0793-26222

David Morgan

Department of Civil Engineering
University of Nottingham
University Park
NOTTINGHAM NG7 2RD

0602-506101

Michael Roberts

Logica (Cambridge) Ltd
104 Hills Road
CAMBRIDGE CB2 1LQ

01-637 9111 Ext 4812

Gideon Senensieb

PA Technology
Cambridge Laboratory
Melbourne
ROYSTON
Herts SG8 6DP

0763-61222

Peter Watts

Logica (Cambridge) Ltd
104 Hills Road
CAMBRIDGE CB2 1LQ

01-637 9111 Ext 4815

Malcolm Williams

Chief Engineer (Advanced Electronics)
Gaydon Technology Ltd
Gaydon Proving Ground
Banbury Road
Lighthorne
WARWICK CV35 OBL

(0926) 644461

Peter Williams

British Rail HQ
Provincial Sector
Room M73
YORK YO1 1HT

(0904) 53022 Ext 2798

University of Leeds

Peter Bonsall	ITS Ext 393
Dr Gustav Clark	Linguistics & Phonetics Ext 6274
Ferri Ghahri-Saremi	ITS Ext 7213
Howard Kirby	ITS Ext 7024
Frank Montgomery	ITS Ext 273
Dr Peter Roach	Linguisitics & Phonetics Ext 6274
Paul Rowlands	Linguistics & Phonetics Ext 6274

E.2 Others interested

Anthony D Allen

Sealink UK Ltd
163-203 Eversholt House
LONDON NW1 1BG

01-387 1234 ext 3910

Fred Atkinson

Lloyds Register of Shipping
71 Fenchurch Street
LONDON EC3M 4BS

01-709-9166

Dr Ryan Beddoes

Plessey 5G
Roke Manor
ROMSEY
Hampshire

(0794) 515222 ext 2402

Mervyn Bramley

Hydraulic & Public Health Eng
Construction Industry Research
& Information Association (CIRIA)
6 Storey's Gate
Westminster
LONDON SW1P 3AU

01-222-8891

J S Bridle

R.S.R.E. M.O.D. P.E. A107
St Andrews Road
MALVERN
Worcs WR14 3PS

Malvern 2733 ext 2890

David Brooke	Consultant (Technical) Information Systems & Technology British Railways Board Room 2 Churnett House Carrington Street DERBY DE1 2ND Derby 42442 ext 3471
Prof Graham R Brooke	Head, Dept of Computer Science University of Hull HULL HU6 7RX (0482) 465696
Paul Collins	Radio & Navigation Dept P161 Building Royal Aircraft Establishment FARNBOROUGH Hants GU14 6TD (0252) 24461 Ext 3428
Prof Frank Fallside	University Engineering Dept Trumpington Street CAMBRIDGE CB2 1PZ
Chris Fowler	M M I Research Unit Dept of Computer Studies Huddersfield Poly Queensgate HUDDERSFIELD HD1 3DH (0484) 22288 Ext 2276/2248
Paul Grant	Imperial College of Science & Technology Department of Geology Royal School of Mines Prince Consort Road LONDON SW7 2BP 01-589-5111 ext 5540 or 5509

Dr Edwin Hancock	Technology Division Rutherford Laboratory Chilton DIDCOT OX11 0QX Oxfordshire (0235) 446292
G R Hobbs	Ferranti Computer Systems Ltd Bracknell Division Western Road BRACKNELL Berkshire RG12 1RA (0344) 483232 ext 3651
Mark Howell	Advanced Electronics Gaydon Technology Ltd Gaydon Proving Ground Banbury Road Lighthorne WARWICK CV35 0BL (0926) 644461
Brigadier Reg King	Military Communications Group School of Electrical Eng RMCS (Cranfield) Shrivenham SWINDON Wilts SN6 8LA 0793-78-2551
Frank Kearney	Construction Eng Research Lab (C.E.R.L.) P O Box 4005 Interstate Research Park 2901 Newmark Drive Champaign Illinois 6180-4005 USA 0101-217 352 6511

Dr Steven Lewis	Scientific Research & Dev Branch Home Office Woodcock Hill Sandridge ST ALBANS Herts A14 9HQ St Albans (0727) 65051
Linda Macaulay	M M I Research Unit Dept of Computer Studies Huddersfield Poly Queensgate HUDDERSFIELD HD1 3DH (0484) 22288 Ext 2276/2248
Christine Mitchell-Brown	North West Water Dawson House Great Sankey WARRINGTON Cheshire WA5 3LU (0925) 724321 ext 33776
Dr Roger Moore	STAG Chairman Speech, Image & Signal Processing Royal Signals & Radar Est St Andrews Road GREAT MALVERN WR14 3PS Worcestershire 06845-2733 ext 2951
Dr Linda Mountain	Dept of Civil Engineering University of Liverpool P O Box 147 LIVERPOOL L69 3BX 051-709-6022
G E Murt	Hydrographic Dept Ministry of Defence TAUNTON TA1 2DN Taunton (0823) 87900

Paul J Neumann — Federal Aviation Administration
Washington DC

202-426-8345

Jeremy Peckham — Logica (Cambridge) Ltd
104 Hills Road
CAMBRIDGE CB2 1LQ

01-637 9111 Ext 4858

D Rix — Balfour Beatty Construction
Bramhall Centre
Ack Lane
Bramhall
STOCKPORT

061-4400700

Mervyn Robertson — John Bradburn (Computing Systems)
St James Mill Road
NORTHAMPTON NN5 5JW

0604 55142

L G Willumsen (Pilo) — Transport Studies Group
University College London
20 Flaxman Terrace
LONDON WC1

01-387-7050

<u>University of Leeds</u>

Andrea Dew — Linguistics & Phonetics Ext 6274
Hakkim Juri — Mechanical Engineering Ext 223
John Uren — Civil Engineering Ext 263

SRREP (3)
Jan 1987