Proceedings of The Institute of Acoustics

DEFINITELY NOT PATTERN MATCHING:- A METHOD IN AUTOMATIC SPEECH RECOGNITION

Julius J. Guzy and Ernest A. Edmonds

Human-Computer Interface Research Unit Loughborough University of Technology

ABSTRACT

A technique for use in automatic speech recognition (ASR) is reported which does not employ traditional pattern matching techniques. The theoretical basis derives from Popper's theory of the growth of scientific knowledge, and claims to solve the problem of applying highly abstract theoretical knowledge to the problem of speech recognition. This problem lies with the construction of bridging relations between theoretical and observation languages. Existing ASR systems attempt to express abstract knowledge in observation language terms. Observation languages are incapable of describing abstract theories. The new technique is a combination of recognition heuristic and research methodology. A detailed description of the technique is presented in the form of a commentary on its application to the detection of the voicing pulse. Starting with a trivial universal theory of the voicing pulse, constraining conditions are derived which exclude all hypotheses which fail to conform to the theory. The tests are implemented as a Prolog program and the results shown. No comparisons are made by the program between prototypical patterns. A new theory is then constructed to take account of the deficiencies in the old one and the process repeated. It is demonstrated that with the growth of knowledge an increasingly greater recognition accuracy is achieved.

PROBLEM

We may consider the problem of speech recognition is as follows. A speaker's vocal tract produces various energies. What we require, is that by performing various measurements upon the acoustic (and possibly other) energies, a computer system should determine whether vocal tract activity is taking place and if so, determine the nature of that activity. Given that a sufficient accuracy is obtained, the system could be used as a Speech Input Interface to a computer application programme. We distinguish between the problem of recognising what a speaker 'says' in the sense of what he articulates, and the problem of interpreting what he says. Note that, from this perspective, recognising what a speaker says is a problem of physics and not of psychology. Thus we are interested in programming a machine to determine the identity of specific physical phenomena as they occur in its environment. A second problem is determining suitable two way communication protocols involving error protection and usability issues. In this paper we shall only be dealing with the problem of determining what speech activity is taking place in the recogniser's environment, specifically, of determining the presence of voicing pulses.We are not concerned with recognising a pattern. We are concerned with recognising specific physical phenomena as they occur in the environment, which is a different problem.

Our approach starts from the premise that every human vocal tract has the same basic design but is slightly different from that of any other, possibly for the reason that it facilitates speaker identification. We take it as fundamental that these differences will be due to relatively minor variations on that basic design. There should, then, exist a number of universal generalisations which may be made about the way in which human vocal tracts interact with their environment such that for any speaker, it is possible to deduce what he is saying, in the sense identified, from the

acoustic energies that are produced. These generalisations may take the form of statements concerning the general laws of vocal tract behaviour. Our work is concerned with the problem of how to discover and use general laws of this kind for the purpose of automatic speech recognition [1,2,3] . As will become apparent in what follows, the method involves using thresholding in the manner of the theory of scientific discovery described by Popper [4,5,6].

A useful consequence of investigating the general laws of vocal tract behaviour is that any statement made regarding those laws will be a universal statement, refutable by a single valid counter example. For example, assume that we have developed a theory descriptive of the behaviour of the voicing pulse. We may have formulated the hypothesis that an energy peak due to the release of the vocal cords, is to be detected at some given time t. We then apply successive tests derived from the theory to attempt to refute the hypothesis. Should the hypothesis be true, according to our theory and it survives the tests or the hypothesis be false according to our theory and the tests succeed in eliminating it, then this fact will corroborate the theory. Should the hypothesis be true and the tests eliminate it, or the hypothesis be false and the tests fail to eliminate it, then this will demonstrate the falsity of our theory. The fact that it is not possible to prove the truth or falsity of any physical theory does not affect our argument. What we will always be looking for will be measurements which constitute tenable criticisms of our theories and therefore force us to invent better ones. Criticisms, in the form of measurements which are contrary to expectation present us with counter examples which we need to explain and take account of in the formulation of an improved theory. It seems intuitively obvious that if we had a true theory, then the procedures employed in the testing of that theory would provide us with the basis of a straightforward and reliable recognition algorithm. For example, in the case of voicing, hypothesize that an energy peak due to a voicing pulse is to be discovered for every time t, and then apply the test procedures. If the theory is correct and the tests are comprehensive then only the valid hypotheses will survive.

Thus, in our approach to the problem of speech recognition, we, firstly, formulate a theory concerning the behaviour of some particular feature of vocal tract activity which we think might be applicable to the construction of useful speech input protocols. We then derive various test procedures with which to test that theory and implement them within a computer program. Next we apply that program to a speech fragment and examine the result. If the result is that only valid hypotheses have survived, then we proceed to apply the program to another speech fragment and so on until either we are satisfied that these procedures could be usefully implemented within a practical recognition system or we find a counter example which forces us to invent a better theory. The key point to note is that we formulate theories in natural language and, we will seek to implement them in ways that enables us to clearly understand the relationship between our ideas and the code. Thus we are able to reflect upon and modify our theories whilst fairly easily updating our software implementation of them. From this point of view, failure is of greater interest than success, since each failure represents an opportunity to develop a theory which is a closer approximation to the truth than its predecessor and therefore brings us one step closer towards realising a practical recognition algorithm.

EXAMPLE

Figure 1 shows a spectrogram of part of the ALVEY speech tape of the utterance "Speech is so familiar a feature of daily life that we rarely pause to define it" [7]. The figure represents about half a second of speech towards the end of the utterance. Only the frequencies roughly in the range 200Hz through 5K Hz are displayed and were used in the analysis. The gray scale values of the image

lie approximately in the range 0 through 63. We will discuss the use of this and other data, obtained from the same source, in order to move towards a theory about striations. The purpose of the discussion is to illustrate the method rather than to make a specific contribution relating to striations.

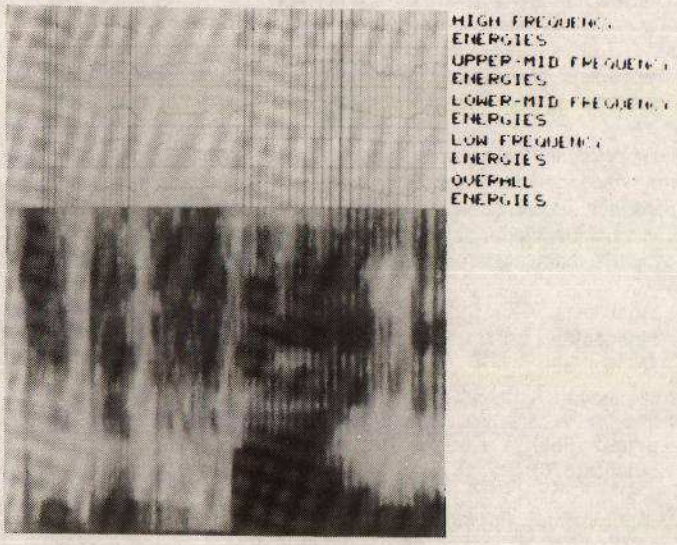


FIGURE 1.

The voicing pulses tend to appear in spectrograms as dark vertical lines and are commonly referred to as 'striations'. Above the spectrogram of figure 1 are some vertical hand drawn lines marking the times at which we think such voicing pulses may have occurred. They merely approximate to the location of those pulses, and are merely used as a guide.

In our example process, a first approximation towards a theory of the physical properties of the voicing pulse is to state the obvious. Namely that a voicing pulse results in a sharp but relatively short pulse of energies produced at the vocal cords. Drawn as a graph of amplitude against time one should therefore observe a peak in the energy values whenever a voicing pulse has occurred. Figure 1 also shows the energy curves obtained by summing and then averaging the gray scale values for each column of pixels in the spectrogram image. Four sets of averages are shown, the first is that of the overall spectrum, the second is of the lower frequencies, the third is of the middle frequencies and the fourth is of the upper frequencies. The curves corroborate the theory.

However, not all the peaks are due to voicing because other phenomena are also occurring. The

energies from those phenomena could be such as to completely obscure those produced by a voicing pulse. Thus any hypothesis regarding the question of whether a peak is due to voicing will always be conjectural. However, one can find reasons for criticising such hypotheses. For example, if we state that we are only interested in detecting voicing pulses produced with a given minimum energy under conditions which will result in the detection of a given minimum energy at the microphone, then one could employ the fact of having observed energies below that minimum value as a strong criticism. Thus if one were to make the hypothesis that a voicing pulse had occurred at some time t and one were to observe that the energies at the microphone at time t were below the critical minimum, then these would be strong grounds for its rejection. Thus, what we have to do in evaluating any voicing pulse hypothesis, is attempt to find similar quantitative reasons for the elimination of that hypothesis. In this example, we may observe that none of the troughs in the curve of the sum of energies shown in figure 1, overlaps the estimated location of a voicing pulse. This might be a reason for rejecting a striation hypothesis. We will therefore make the provisional assertion that no striation pulse may be inferred in cases where there is a trough in the sum of energies curve.

To implement the above method, a Prolog program was written which initially, generated a striation hypothesis for every time t under consideration. This was the simplest possible starting point and amounted to nothing more than the creation of a set of entries in a database where each entry has the form 'striation(T)' and T is a unique integer denoting an instant in time. Since our digitised spectrogram was a 512 by 670 matrix, this resulted in 670 such entries, viz:

```
striation(1).
striation(2).
striation(3).
    .
    .
striation(669).
striation(670).
```

The idea is to now perform test measurements on the basis of which we may remove entries from this database. Thus, for example, if we perform a measurement which indicates that the energies fall below a given minimum at time t=2 then we remove the entry: striation(2) from the database. After all tests have been performed, then those entries remaining in the database represent moments in time for which no grounds could be found for the rejection of a striation hypothesis.

In what follows we shall be describing the formulation of various theories which will be composed of what may here be called component theoretical ideas from which the refutation tests will be derived. To facilitate exposition we will refer to these theoretical ideas by reference to the tests derived. Thus for example, the theoretical idea to be described next will result in a test which we shall refer to as 'refutation 1', a reference which we will employ to likewise refer to the theoretical idea from which it stemmed. The state of the entire theory developed up to a given point in time may therefore be defined by reference to the ordered list of the numbers of the tests employed. For example, we will initially be forming a theory which employs tests with labels: refutation 1, refutation 2, ..., refutation 5. The state of the theory may then be defined as the set T(1,2,3,4,5). As our ideas evolve the identifying numbers of new tests will be added to this set and likewise as ideas are abandoned so the appropriate numbers will be deleted. Thus, if for example, refutation tests 2 and 4 are abandoned and a new test with identifier 6 is added, then the new state of the theory may be defined as being T(1,3,5,6). Effectively, what this is expressing is the fact that with each iteration

we are trying to explain the same set of observations as the previous theory explained, together with those which it failed to explain, i.e. those on which the tests derived from the old theory, failed.

Following through our previous observation that no voicing pulse coincided with a trough in the sum of energies curve, we sought out the energy maxima and minima and deleted all database entries where there existed a minimum at time t. Thus for example, a minimum was found at time t=20 and the entry: striation(20) removed from the database. The results of this test (refutation 1) for the whole spectrogram are presented diagrammatically in the top row of figure 2. Here each of the blocks composing the row, represents either one or a contiguous set of surviving voice pulse hypotheses. No valid hypotheses were rejected by this test. However a large number of quite obviously invalid hypotheses remain.

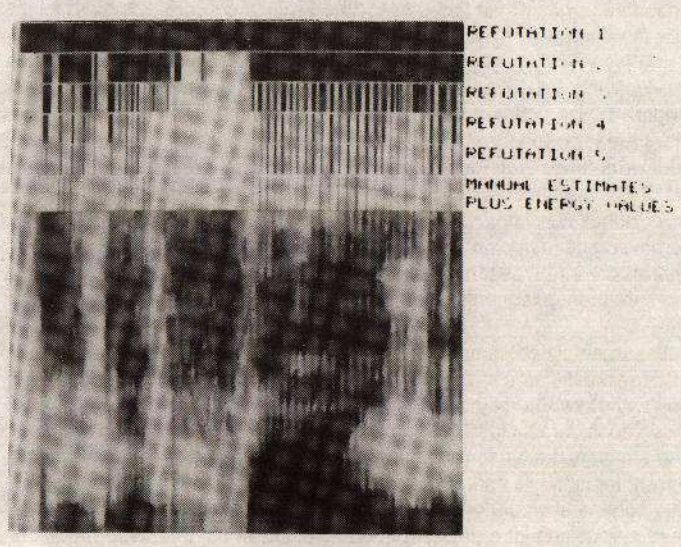


FIGURE 2.

One property of the voicing pulse as it appears in spectrograms of speech spoken directly into the analyser microphone is that it is always accompanied by a dark region in the lower frequencies commonly referred to as the 'voice bar'. Hence we may eliminate a striation hypotheses for any time t at which the energies in the lower frequencies fall below a given threshold. The results of this test: refutation 2 are shown as the second row of results in figure 2. Again, no valid hypotheses were eliminated.

One feature of the energy pulses associated with striations as they may be observed in broad band

spectrograms is that they appear to form long vertical ridges. However, notice that strictly speaking this assumption is incorrect as we discovered when analysing the spectrogram shown in figures 7 and 8 (c.f. below). One thing which might be assumed never to occur except in the presence of other energies, is that the striation be bounded on either side by higher intensity values over any reasonably long contiguous band of frequencies. This then constitutes the third test criterion: refutation 3. The results are presented in figure 2 and again, no valid hypotheses were rejected.

A corollary of the preceding hypothesis is that when a voicing pulse has occurred, then in the absence of other energies, it should be possible to observe at least one ridge like feature. That is, there should exist at least one contiguous frequency band of energies of a minimum intensity which is bounded on either side by lower intensity energies. This test: refutation 4, was implemented and the results are again shown in figure 2.

It may be seen that after refutation 4, with two exceptions, all of the locations of the striations have been roughly approximated by the hypotheses with clear gaps in the hypothesis sequences existing between them. A number of surviving hypotheses refer to two pulses which as far as we are aware are due to vocal tract phenomena other than those associated with vocal cord activity. It might be possible to further reduce the number of hypotheses by eliminating those which referred to those parts of an energy pulse during which the energies were building up to a peak value and those during which the energies were decaying. To do so we implemented a condition which looked for a broad band of contiguous values (approx 1kHz) in the frequency dimension for each time t which were lower than those in the same frequency band for any of the preceding time slices t-1, t-2, t-3, or, which over that same bandwidth were lower than those in the corresponding three succeeding time slices, t+1, t+2, t+3. The results of this test: refutation 5, shown in figure 2 were surprisingly successful. They not only dramatically improved the overall accuracy of the estimated positions of the striations but also removed all the hypotheses referring to the two non-voiced pulses. For the purposes of this demonstration we felt reasonably satisfied with the results so far obtained and decided to test them upon another speech fragment.

A very noisy portion of the same speech sample was selected for this purpose and is shown in figure 3 together with our hand drawn estimates of the location of the striations and the energy levels in the different frequency bands as before. As may be seen from the close proximity of the striations, the voice at this point had risen considerably in intonation. It is immediately obvious from the lack of temporal resolution of the striations that in forming our theory we had failed to consider the rate of pulsing. Applying the current tests to this speech segment should therefore be expected to fail. There seemed indeed to be little point in even performing the tests in order to merely look for success or failure. However, since we are looking for a greater understanding, we felt that other problems might be revealed and therefore proceeded nevertheless.

As expected, many of the results, especially for the latter half of the speech fragment are valueless. Furthermore, the differences in overall energy values are so small that even though in the first part of the fragment they appear to provide a valid result, any justification for eliminating hypotheses merely because they happen to lie in an energy dip, would seem tenuous. The fault in theory T(1,2,3,4,5) was that we had, amongst other things, failed to consider the existence of energies due to non-voicing phenomena. The test: refutation 1 was not valid. We decided not to employ it further. The current state of the theory therefore became T(2,3,4,5). By discarding this test, the constraints on the survival of the hypotheses have been relaxed. Thus, no hypotheses that had survived the tests will be refuted as a result. This has the useful consequence, that no matter how

DEFINITELY NOT PATTERN MATCHING:- A METHOD IN AUTOMATIC SPEECH RECOGNITION

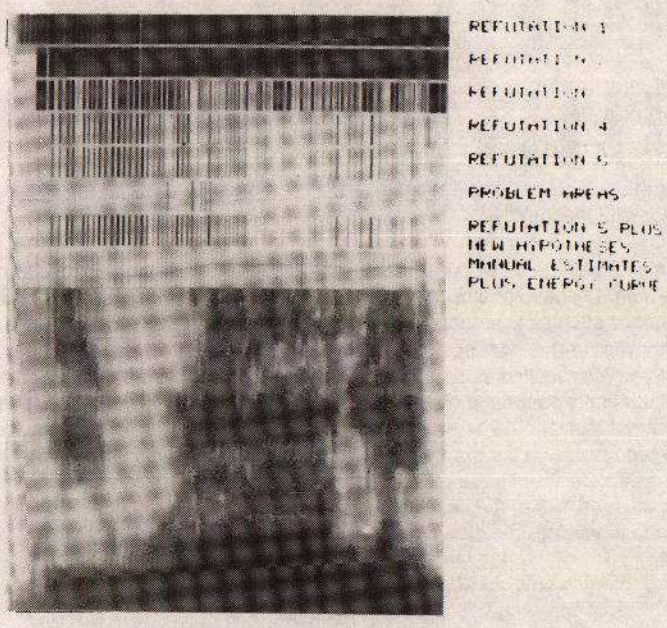


FIGURE 3.

complex our theories and related test procedures may become, the result of deleting a test from the set, will only have the effect of permitting the survival of those hypotheses whose refutation is dependent on its existence. By corollary, the introduction of a new test will also have a determinable effect on the survival of the hypotheses. Thus, the overall system has a modularity which facilitates easy modification.

Examining the results of applying the remaining tests, showed that in the first half of the speech fragment, a hypothesis survived to identify all but three of the voicing pulses present. One non-voicing pulse was also identified. Ignoring for the moment the question of how to deal with that non-voicing pulse, it seemed to us to be of some interest to consider how one might deal with the problem of missing pulse hypotheses in a context such as this. We therefore decided to explore this idea independently of the task of deriving a unified striations theory. We will describe this detour in our thinking because it was to have an influence on the evolution of the theory as a whole.

A striking fact about these incorrectly deleted hypotheses is that their location seems intuitively obvious. For example, if one looks at the diagram below and imagines the top row to represent the location of some striation hypotheses then it seems obvious that two hypotheses are missing and that the true picture should look something like the bottom row, (missing hypotheses shown in gray).

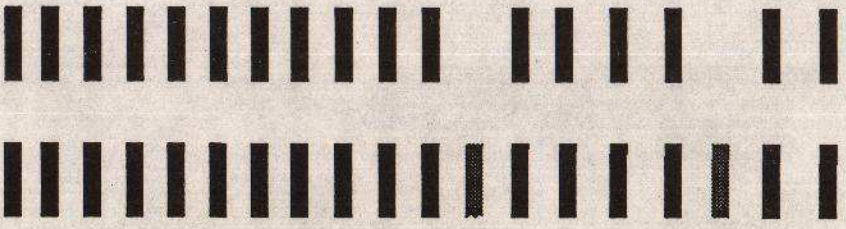


Diagram 1. (see text for explanation)

The reason why one might feel that an error had occurred here is because our understanding of the way in which the vocal cords behave seems to preclude the possibility of individual pulses failing to be produced in a sequence such as this one. (Since then we have learnt that indeed such events can occur but that they are quite rare). Under such an assumption one may therefore conclude that either the refutation test which resulted in the elimination of the striation hypothesis produced an incorrect result or that the entire sequence of apparent pulses is due to some phenomenon other than that of voicing. An incorrect result could have arisen for a variety of reasons, for example, that it was not a particularly good test, that the theory from which it was derived was wrong, that energies due to other phenomena interfered with the test, and so forth. Hence one arrives at a point of conflict between opposing hypotheses. Either the decision to refute the hypotheses in question was incorrect or the pulses are not due to voicing.

The possible occurrence of such an error seems amenable to detection using formal methods. Between each time contiguous block of surviving hypotheses there exists a gap which in the case of the error locations appears to be roughly twice as long as any of the others. Hence one appears to have here the basis for an algorithm with which to detect the occurrence of problem areas such as the above. We therefore wrote a program to detect these problem areas. The problem locations identified are shown in figure 3 (problem areas) . Given that the problem areas have been identified, the question arises as to what may be done about them. One reason for suspecting that a refutation test might have resulted in an incorrect decision would be if a sudden increase in energies had occurred, or the vocal tract was rapidly changing its configuration. For example, in this case both reasons seem to be operative, first the energy increase due to frication and secondly both an increase in pitch of voicing and an upward movement of the second formant. Under such conditions as increases in pitch one might suspect that the muscular movement required to obtain it was less than smooth. Hence if such tests were to be performed and the decision obtained that a refutation test had wrongly succeeded, new voice pulse hypotheses could be generated for the problem times identified as shown in the row labeled: 'refutation 5 plus new hypothesis' in figure 3. It seems to us however, that such an approach could only be properly developed given our having first acquired a far greater understanding of the behaviour of striations than we have at present.

Returning to the main task of further developing our theory, we increased the temporal resolution of our data to a level which took account of the high vibration rates of the vocal cords are capable. This effectively meant increasing the resolution until the individual striations in the problem area became clearly visible. We decided for the pragmatic reasons of processing time and memory storage, to temporarily ignore the energies present outside of the range 0.5 through 2.5 KHz. Figure 4 shows a close up of the problem area smoothed in the frequency domain. A number of minor

modifications to the test measurements seemed appropriate to take account of the higher level of resolution. These modifications do not of course imply any change to the underlying theory. Because of the frequency ranges being considered, we could not apply the test: refutation 2. However, for the time spanned by the spectrogram, the test had already failed to refute any hypotheses and therefore did not actually need to be repeated. For reasons given previously, the test: refutation 1, based on dips in overall energy values was not employed either.

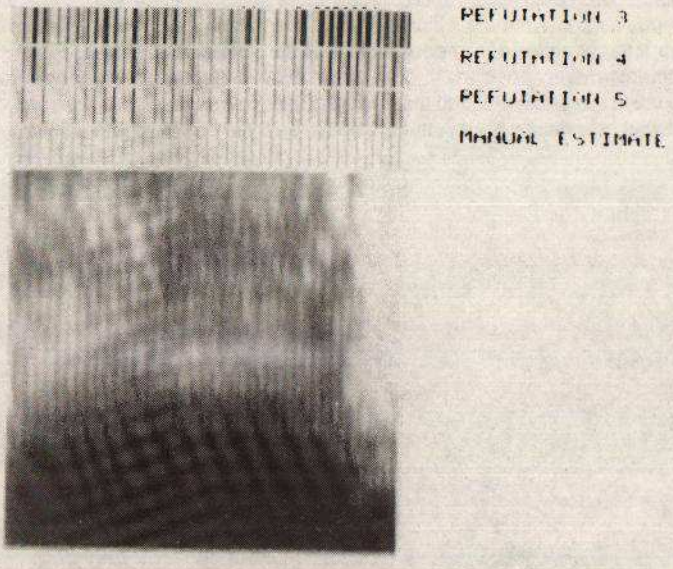


FIGURE 4.

The results of the tests, as shown in figure 4 are clearly unsatisfactory because they fail to eliminate a sufficient number of false hypotheses and in places appear to have given a result which with respect to our estimation of the location of the voicing pulses seemed dubious. The reason why this should have happened seemed due to the presence of energies due to phenomena other than voicing. We could at this point have put more effort into discovering the salient characteristics of the individual voicing pulses. However, a more urgent issue seemed to be that of how to deal with this type of situation which may be sure to occur occasionally no matter what tests are devised. This was an issue which we had already explored as described above.

As mentioned, the reason for the unsatisfactory results was clearly the presence of energies due to frication. Nevertheless, to the human eye the striations are quite clearly visible. The reason seems to be that the turbulence energies seem to come in bursts at quite distinct frequencies, such that in

the other frequencies the striation pulses may be easily discerned by the fact of their regularity. Thus what seems necessary is to locate frequency bands which appear to contain regular pulsations and then employ the presumed regularity to predict the location of voicing pulses. These predicted pulses could then be used as a basis for refuting those survivors of the previous tests which fail to conform with the discovered regularity. This of course is related to the idea explored previously.

As an unsophisticated first step, the frequency domain was divided up into ten contiguous equal sized frequency bands (approximately 200 Hz). Next, for every time t, it was hypothesized that t marked the midpoint of a voicing pulse one either side of which, separated by a fixed time interval, existed two other voicing pulses. Thus one was hypothesizing a sequences of five regular pulses. For each time t a number of such hypotheses was made, each for a different pulse interval. Because of the width of the striations at this level of resolution it was thought appropriate (incorrectly as was later revealed) to consider each pulse peak to be roughly three spectral slices wide. Tests were then performed on each hypothesis to eliminate it if there was a great variation in the intensities of the hypothesized peaks or if the energies of any peak were less than or equal to the energies of any one of the hypothesized intervening valley positions. Figure 5 shows these sequences superimposed over the spectrogram at locations approximating to the position of the frequency bands for which they were hypothesized. It was now necessary to attempt to join the hypotheses together wherever

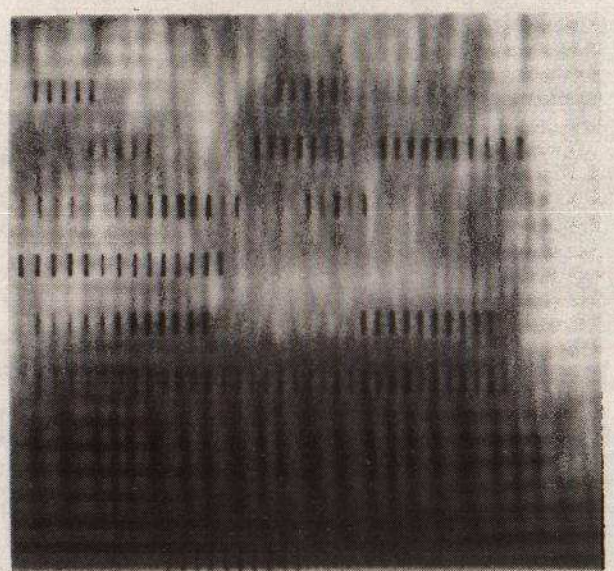


FIGURE 5.

possible. The first step was to merge where possible those hypotheses which had the same time interval. This involved first merging those which overlapped for particular time sequences. Thus for example, if one had a sequence of hypotheses for pulses at times [10,15,20,25,30] and [5,10,15,20,25] these gave the sequence [5,10,15,20,25,30]. However, one also frequently obtained the case where though the interval was the same the times of the pulse midpoints differed by just one. For example, the sequences [10,15,20,25,30] and [11,16,21,26,31]. These clearly were referring to the same pulse, and were merged to form blocks of sequences

[[10,11],[15,16],[20,21],]25,26],[30,31]].

When merged with the earlier string this gives the result

[[5],[10,11],[15,16],[20,21],[25,26],[30,31]].

Next, pulse sequences which differed in interval but which overlapped over contiguous blocks were merged wherever possible. For example, the above sequence and the sequence [[14],[20],[26],[32],[38]] would be merged to give the result

[[5],[10,11],[14,15,16],[20,21],[30,31,32],[38]].

Following this there remained quite a number of hypothesized sequences which could not be merged together in this way. Their existence and the points at which the merge failed were indicative of problem areas, since such an eventuality could only be explained by for example there being two or more speakers speaking simultaneously, or the existence of a pronounced echoing pulse as sometimes occurs when the vocal cords snap back together, or to one or the other or both of the conflicting sequences being an incorrect hypothesis.

Four categories of problem were identified. The first being the case where a merge could not be performed between adjacent sequences having the same interval,

e.g. [[31],[36],[41],[46,47],[51,52],[57,58],[63,64],[70],[76]]

and [[56],[62],[68],[73],[78]],

the problem existing between time 68 and 78.

The second type was that of overlaps of the form

[[30],[36],[42],[48],[54],[60],[66]]

and [[38],[44],[50],[56],[62]]

where the problem exists between times 36 and 66. The third was of the same form but for different intervals between pulses. The fourth was that where the head and tail of sequences of different intervals partly overlapped over a contiguous set of blocks but failed on others. For example

[[70],[75],[80],[85],[90]]

and [[73],[79],[85],[91]]

where the problem exists for times 70 through 75.

The method was to create a pulse hypothesis for all times t at which a pulse was hypothesized by a pulse sequence hypothesis. Call these 's' hypotheses to differentiate them from the first set of hypotheses to be tested. Call those the 'f' hypotheses Thus for the sequence [[10],[15],[20],[25],[30]] the s hypotheses would be 10,15,20,25,30. The idea now was first to remove such of these as could be deleted by reference to the problem areas and then use the survivors in order to refute the f hypotheses. We assume this approach to hold only in cases where we have sequences of at least 20 contiguous pulses, since for shorter sequences the justification for the method seems less intuitively obvious. Thus, conflicts of the second type were resolved without further testing in favour of the longer hypothesis, i.e. all the s hypotheses for the discounted sequence would be deleted. No problems of this second type existed for other than very short sequences. In the case of problems of the first, third and fourth type, if no f hypotheses existed for one of the problem times t then the corresponding s hypothesis was deleted. The remaining s

hypotheses were then examined and wherever a time t existed such that no s hypothesis was present and there existed an s hypothesis at times t-1 and t+1, then on the grounds of an earlier argument, an s hypothesis was generated for time t. The start and end times of the period covered by the s hypotheses was then noted. The f hypotheses which fell within this time interval were examined and those which did not overlap with an s hypothesis were deleted. Again gaps of one time slice in duration were filled by generating a hypothesis for those times. The results of this test: refutation 6, and all the other tests are shown as figure 6.

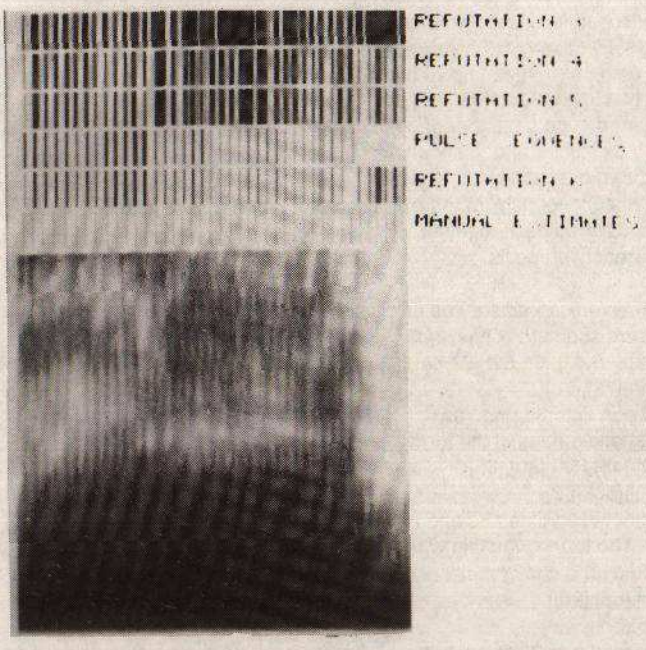


FIGURE 6.

In order to test the theory, another portion of the spectrogram was selected, again looking for a region where the method might be expected to fail. The selected region together with the results is shown as figure 7. One of the reasons why the approach failed was because of the assumption concerning the width of a striation peak and valley (c.f. above). We changed this part of the program so that only a single time slice was considered and re-ran it. The improved results are shown as figure 8. Notice that although only one valid hypothesis appears to have been refuted, there nevertheless remains an unacceptable number of invalid hypotheses which survived refutation. The reason for this appears to be in part the fact that contrary to our theory regarding the simultaneity of the energies of a voicing pulse (c.f. above), there exists a considerable divergence between the time at which energies at different frequencies attained their peak values! Thus, the tests based on the notion of a simultaneous rise in energies were shown to be invalid. Furthermore, the

'distortions' can cause the pulse sequences hypothesized within one 200 Hz frequency bandwidth to completely overlap the interval between the pulses hypothesized within another of those bandwidths. What seems to be happening is that with frication, energies in some of the frequencies are being absorbed or in some way delayed in their passage into the environment. In effect what this would all seem to indicate is a need for a revision of the simultaneous pulse of energy theory and a less ad hoc approach to seeking the pulse sequences than the one we have pursued so far. The process must then continue, but it is hoped that sufficient has been described for the method to be understood.

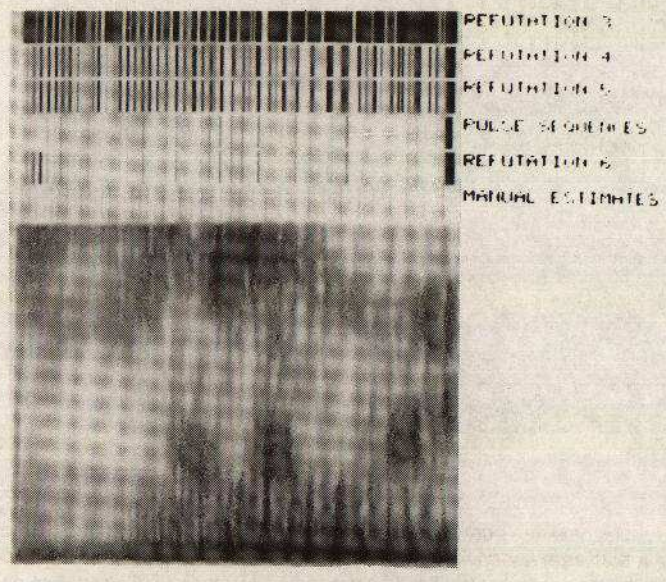


FIGURE 7.

COMPARISON

We will now briefly discuss the relationship between the method described above and pattern matching. The term pattern matching is used here to mean that family of recognition techniques which involve the performance of some kind of similarity comparison between a set of input data and a set of prototypical exemplars. The goal is to assign each item of input data to that exemplar to which it bears the greatest similarity [e.g. 8]. Within this family of pattern matching approaches must also be included any parsing mechanism which involves the use of an explicit representation to which the resultant analysis of the input data must then be matched.

The refutation approach is explicitly concerned with determining the identity of given phenomena in

the recogniser's environment rather than with the detection of similarities between patterns. Pattern matching, by contrast, is explicitly concerned with determining the identity of that set of reference data to which the input bears the greatest similarity.

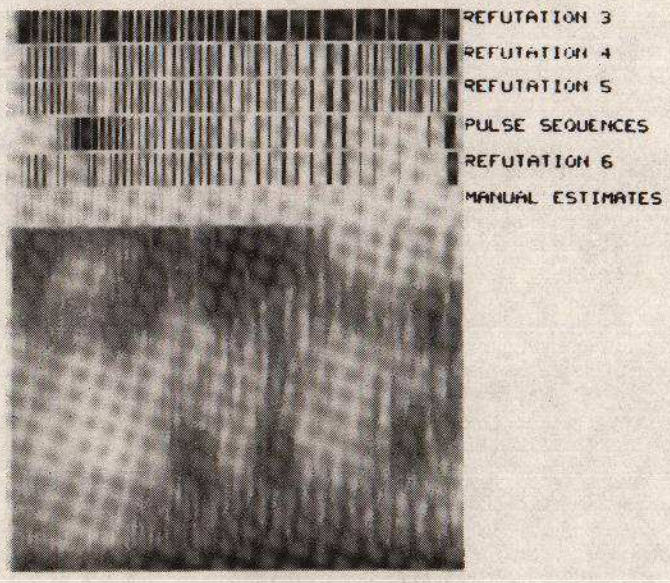


FIGURE 8.

In spite of the great differences between human vocal tracts at the level of absolute measurement, there nevertheless exists a dominant underlying design which unites them. Hence the way in which one vocal tract interacts with its environment will exhibit properties common to all vocal tracts. These properties exist only at the highly abstract-theoretic level of human understanding. One can look upon the speaker-normalisation procedures employed by pattern matching as an attempt to express this abstraction. However, the expression of abstract-theoretic conceptualisations in such terms presents considerable difficulty and may be impossible [9]. The refutation approach described above takes the abstract-theoretic conceptualisation as its starting point. Therefore, it is possible to make use of any testable physical theories, relevant to the problem, in order to construct the complete theory postulated.

In the refutation approach we are not proposing a simplistic reversal of technique. We are not making the suggestion that instead of looking for positive, corroborative, instances of a characteristic of a phenomenon, we should look for negative, refuting, instances. To do so would lead inexorably back to a description of phenomena in terms of measurement. Rather, what the refutation

approach does, is to use the principles of the theory of a phenomenon to derive tests which in the case of examining an actual instance of that phenomenon occurring in the environment, gradually enables us to build up a picture of the characteristics of this singular instance. That description will be produced from the viewpoint of the theory. That is, the refutation approach enables us to perform a constructive analysis of the phenomenon in question. A very simple example of this was given above in the case of the third spectrogram, where through the assumption of the regularity of the voicing pulse, it became possible to ignore the irrelevant and predict that which was not directly observable.

It is interesting to compare pattern matching and the refutation approaches from the standpoint of what may be learnt by their pursuit. The refutation approach is dedicated to the determination of the universal characteristics underlying the behaviour of given phenomena. Every time that a recognition failure occurs, the theory or the test procedures need to be reformulated. Assuming that our inventiveness does not fail short of the task, this means that in principle, one should be edging ever closer to the truth. In pattern matching that which may be learnt is constrained to being the parameter set and discrimination criteria appropriate to a given set of target pattern classes. Any substitution of a member of that set or addition to to that set may require a different parameter set to be employed. Since one cannot enumerate all the combinations of target classes that are possible, one can never arrive at a parameter set and set of discrimination criteria which one may be certain will work in all cases. The problem of certainty is common to both approaches. The crucial difference is that under the refutation approach the circumstances under which the recognition algorithm will provide a particular result is given by the the tests employed. Under pattern matching however, there seem to be no grounds for predicting that subset of the environmental phenomena for which a given parameter set and set of discrimination criteria will produce a given result, short of referring to those combinations which have already been tried.

It is interesting to note that a theoretical difficulty exists with respect to the evaluation of ASR system based on pattern matching. The reason is that generally one is not comparing like with like. A system which works better that another on one vocabulary, might work worse than the other for a different vocabulary or speaker. The root of this problem lies precisely with the lack of universality inherent to pattern matching. In the case of the refutation approach, because any system based on this approach will be laying claim to the universality of its theories, the contrary is true. The system may be tested for any speaker under any acoustic conditions. Where systems only have one or two recognition phenomena in common, then the comparison of performance can be constrained to just those phenomena. Failure is instant and evaluation is a function of the generality of the test criteria. Thus, in principle, anyone can test both the underlying theories and the performance of the system according to any tests that he may care to devise. These criticisms would then need to be answered by the formulation of better theories or more extensive test criteria.

DISCUSSION: SOME ISSUES FOR SPEECH RECOGNITION

The approach that we have presented is trying to deal with the problem of what has been called the 'primacy of the abstract' [9]. The hypothesis of the primacy of the abstract is that in our discovery of the world we do not move from the particular to the abstraction, but rather from the abstraction to the particular. We observe things as being similar because they satisfy the criteria of the unifying abstraction. Dissimilarities are those respects in which an object or idea fails to satisfy the criteria of the abstraction. It is this which makes it possible to speak of some measurements as being

irrelevant. They are irrelevant because they do not belong to a dimension in what has been called 'pattern space' [e.g. 8], over which the abstraction defines any test criteria or because a measurement in that dimension is inapplicable to the object or idea in question. Thus similarity between objects or ideas will always be "in some respect", and objects which are not comparable in one respect might be identical in another respect depending on the abstraction.

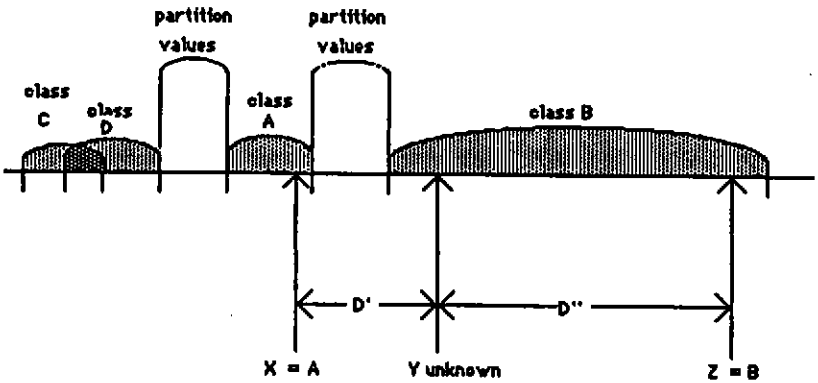
If we consider just one dimension of measurement, then measurements which lie outside the boundaries defined by an abstraction over that dimension can be said to be relatively closer or further from those boundaries. Thus, for any object or idea not belonging to the abstraction but which may be measured in one of the relevant dimensions, one can state how close it is to the boundary conditions. One may see immediately that questions of similarity can be asked in two distinct ways. First is that of asking whether two objects or ideas fall within the criteria of acceptability of a given abstraction, i.e. that they are similar in the given respect. Second is that of asking of a given object, the extent by which it has failed to satisfy the criteria of acceptability defined in a designated dimension.

If we are interested in determining the nature of an abstraction, for example, the general laws describing the behaviour of a given phenomenon, then we need first to determine at least one of the dimensions relevant to that abstraction. Given that we have correctly identified such a dimension, the next objective has to be that of determining the constraints imposed by that abstraction on the class membership criteria in that dimension. To do this we will need to guess at specific values and we will be interested in finding how far each successive approximation is from the class membership boundary conditions. What we shall want to do is to stay just outside of those boundary conditions whilst at the same time edging ever closer to them. We can determine whenever our approximations are on the wrong side by the fact that this results in the elimination of a valid member of the class. Given a seemingly satisfactory approximation one may then try to determine the identity of the next dimension. It is the demonstration of this process which formed the purpose of our example.

Alternatively, one may just be interested in discriminating between some finite set of classes. Here the problem is first to find a dimension common to them all and secondly try and discover whether the classes differ from each other in that dimension with respect to their abstractions. For those that do, a simple discriminant value would suffice to define the measurements necessary to accomplish a partition over the set of classes. Important to notice is that the question of the proximity of that value to the actual boundary conditions is immaterial so long as it lies between the boundary conditions of the classes to be partitioned (e.g. diagram 2).

Consider the use of objects which are members of these classes as a means of determining the discriminant value. For any two classes A, B, say, their boundary conditions need not be equal with respect to the degree of constraint that they place on the class members. It may also be the case that the range of the boundary conditions of any two or more classes overlap as shown in the diagram.

DEFINITELY NOT PATTERN MATCHING:- A METHOD IN AUTOMATIC SPEECH RECOGNITION



X = A indicates measurement of object X of class A
Z = B indicates measurement of object Z of class B
Y unknown indicates measurement of object Y
result gives distance D' < D'' therefore Y is incorrectly assigned to class A

DIAGRAM 2: Why simple pattern matching techniques in the ignorance of class boundary conditions cannot guarantee a correct classification decision.

Merely by looking at the diagram one can see that tests of proximity to the values of the class exemplars will not guarantee a successful partitioning of the classes in that dimension. It follows that simple tests of distance in ignorance of the actual boundaries of the class definition criteria, must in general be considered unlikely to succeed in their purpose. A rational approach to attaining a successful discrimination would seem to indicate a need to determine those values between which lie the definition criteria of each class and that would seem to necessitate the formation and testing of universal theories. If our arguments are correct, then traditional approaches to the problem of speech recognition are theoretically unlikely to succeed in delivering reliable results. The approach which we have described in this paper, represents an attempt to find a viable alternative.

REFERENCES

[1] CONNOLLY, J. H., EDMONDS, E. A., GUZY, J. J., JOHNSON, S. R., & WOODCOCK, A., 'Automatic speech recognition based on spectrogram reading', International Journal of Man-Machine Studies, 24, pp. 611-621(1986).

[2] GUZY, J. J. 'The acquisition of linguistic knowledge from visible speech spectrograms: a proposal', International Journal of Man-Machine Studies , 16, 327-332. (1982).

[3] GUZY, J. J., CONNOLLY, J. H., EDMONDS, E. A., & HASHIM, A. A. 'A feasibility study into a system for direct speech input to computers. Final Report', SERC Grant No.Gr/B/54599, Leicester Polytechnic. (1980).

[4] POPPER, K. R. 'Objective knowledge', Oxford University press, Oxford. (1979).

[5] POPPER, K. R. 'The logic of scientific discovery', Hutchinson, London.(1980).

[6] POPPER, K. R. 'Realism and the aim of science', Hutchinson, London, (1983).

[7] Alvey speech tape: Digital and acoustic tapes of a speech fragment (including laryngograph signal) prepared by the Alvey Speech & Natural Language Club for a workshop held at University college London, 15 September 1986.

[8] LEVINSON, S. E., 'A unified theory of composite pattern analysis for automatic speech recognition', In Fallside, F. & Woods, W. A., Eds., Computer speech processing, Prentice Hall International, London. (1985).

[9] CARNAP, R. , 'The methodological character of theoretical concepts'. In Feigl, H., & Scriven, M., Eds., Minnesota Studies in the Philosophy of Science, Vol I, University of Minnesota Press, Minneapolis. (1956).

[10] HAYEK, F. A., 'The primacy of the abstract', In Koestler, A., & Smythies, J. R., Eds., Beyond reductionism, Hutchinson, London. (1969).

Proceedings of The Institute of Acoustics

THE PROBLEM OF CAPTURING LINGUISTIC AND PHONETIC KNOWLEDGE

Marcel A.A. Tatham

Centre for Cognitive Science, University of Essex

Introduction

There can be no doubt that the success of text-to-speech synthesis systems and latterly the improvements in automatic speech recognition systems owe much to the bringing in of information from linguistics and phonetics. But it seems that a mistake has been made in what has been brought in. This is particularly true in synthetic speech where the ideas have had longer to become entrenched. The speech the latest devices produce is just not good enough. The failure has little to do with the design of the actual synthesiser. Whatever faults synthesiser hardware may have they are patently not the limiting factor. So in speech synthesis by rule it is easy to reach the conclusion: it must be the rules. It is not - at least not in the sense that if we devote more effort to refining the rules all will come right. What is wrong is that the *whole system* (I use the word in the singular since all the leading systems - JSRU and its derivative BTalk, MITalk and its derivatives DECtalk and Prose-2000 - are all similar in type) has been ill-conceived. The fault lies as much with the linguists as it does with the engineers. The same is true of automatic speech recognition systems which employ grammars in a top-down effort to disambiguate the output of low-level bottom-up analysers.

Linguistics

Linguistics is the science which deals with grammars. The theory is exemplified in models which have an exhaustive formal structure which is utterly explicit [1]. In no sense can the argument that linguistics is informal or inexplicit be sustained. The metatheory is equally clear: linguists know what they are doing and why they are doing it. Like any developing science linguistics has known its infighting, but the general principles and goals remain constant and, from our point of view, the main points are established and unarguable.

The theory of linguistics is descriptive and has explanatory aims. It describes the knowledge a person has of his language [2]. It is this which enables him to encode his thoughts for transmission to another person and to decode the thoughts transmitted to him by another person. The carrier which concerns us is sound, though of course there are others. In addition and in common with other branches of cognitive science linguistics seeks through its explanatory power to throw light on the structure and workings of the mind. This latter goal does not for the moment concern us here.

What has been misunderstood in bringing linguistics to bear on speech synthesis and automatic speech recognition research is just what it is that linguistics describes. It characterises the *knowledge base* human beings have which they access for the purposes of the encoding and decoding procedure. Linguistics has something to say about the structure of the knowledge base and the system constraints placed on it. But in general it has nothing to say about accessing procedures or the encoding and decoding algorithms. Although the knowledge base description may contain rules which delete, add or transform elements they are in no sense intended to be interpreted as elements in some specific encoding or decoding algorithm. It is only in the theoretical sense that the entire knowledge base describes the potential of the entire language, but it is strictly not true that some part of the knowledge base constitutes an algorithm. What a part of the knowledge base might constitute is a description of those rules which might be

LINGUISTIC AND PHONETIC KNOWLEDGE

accessed by some un-prescribed algorithm for the purposes of some encoding or decoding procedure to link a particular concept with a particular soundwave. This is the source of the error researchers in speech synthesis and automatic speech recognition, engineers and linguists alike, have made.

Fig.1

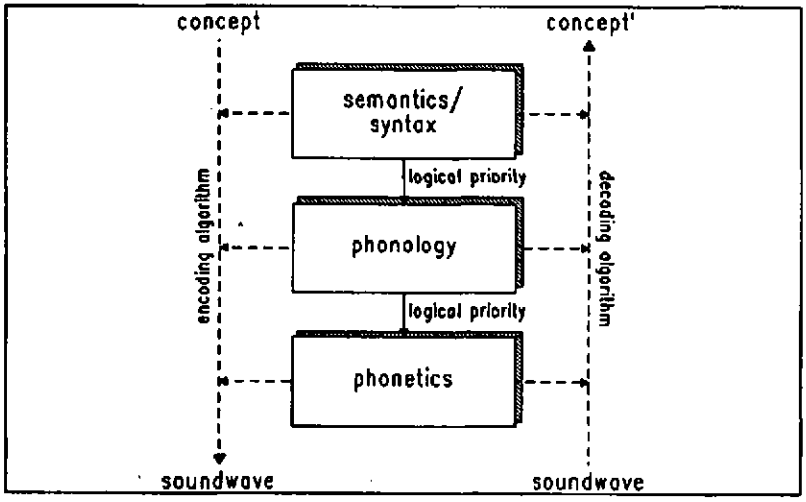


Figure 1 illustrates the general model in linguistics. The overall knowledge base is subdivided into different components. There are formal and substantive reasons for the subdivision which need not concern us here. Let us simply refer to these components as the semantic/syntactic, the phonological and the phonetic knowledge bases. The vertical line linking the components indicates that particular knowledge bases are logically prior to others: they are not temporally prior or procedurally prior since linguistics has nothing to say about timing or procedural activity. The dashed lines are not part of linguistic description. They indicate potential procedural and accessing flow in some system outside the domain of linguistics, and are there to show a relationship between the knowledge bases other than the logical one dealt with by linguistics.

The semantic/syntactic and phonological knowledge bases each contain

(a) lists of the primitives associated with their particular level in the grammar, and

(b) sets of rules constraining the co-occurrence of these primitives.

So at the phonological level [3] the knowledge base contains information on

(i) the set of phonological features in use in the language and the rules constraining their combination in the formation of phonemic segments in the language;

(ii) the set of phonemic segments actually available in the language and the rules constraining their sequencing in the formation of words;

(iii) rules characterising transformations of the phonemic segments in particular contexts with other phonemic segments;

(iv) a prosodic sub-section comprising primitives and rules for the assignment of prosodic contours to words and word groupings.

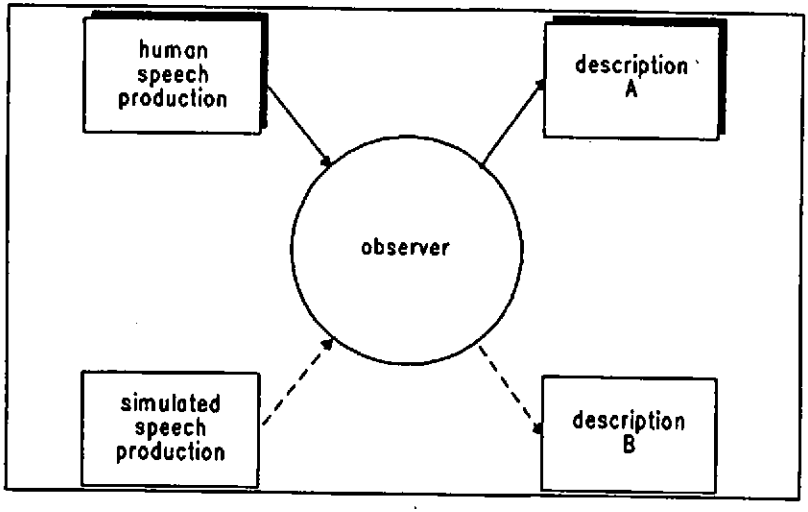
Together these primitives and constraints ultimately characterise the extrinsic allophonic patterning used to encode all sentences in the language. It is important to note that it is all sentences, not a particular sentence.

At the phonetic level things are slightly different. Until recently it was believed that there was little in phonetics of interest to linguistics, since apart from a logical entry point accessed by strings of extrinsic allophones the entire component was dominated by constraints of myodynamics, aerodynamics and acoustics. While this was believed to be the case the level was outside the domain of linguistics, since linguistics is a cognitive science and such physical phenomena were anything but cognitive. But we are now coming to believe [4] that although there are many physical constraints on the production of speech sounds it is nevertheless the case that these constraints can be systematically inhibited or enhanced, and indeed are fine tuned under cognitive control for linguistic purposes. The ability to manipulate physical constraints under cognitive control is extremely important to us when we come to consider variability in speech. If cognitive control of physical constraints is possible it must be the case that the nature of those physical constraints must be known to the system. Hence the need for a phonetic knowledge base ennumerating those constraints as part of general linguistic knowledge.

It seemed necessary to go into some detail about what linguistics can tell us and what it does not. To repeat: linguistics is a descriptive characterisation of the knowledge a human being has to enable him to encode and decode speech. It says nothing about the acts of encoding or decoding. In speech synthesis and automatic speech recognition on the contrary the focus of attention is precisely on encoding and decoding. Synthesis and recognition are *not* equivalent to descriptive models: they are simulations.

Description and simulation

Fig.2



Descriptions and simulations are very different objects and it is important not to confuse the

two. Fig.2 illustrates their relationship. An object, human speech production, is observed by the scientist who produces a description of it. This is description A in the diagram, and is equivalent to the description provided by linguistics together with a characterisation of the general algorithm and procedures for accessing the knowledge bases. A second object, simulated speech production, is also observed by the scientist. He produces description B of the simulated speech. The more like human speech production the simulation becomes so description B approaches description A. Our criterion of success in simulations is the degree of difference between descriptions A and B. The point of this diagram is to indicate that description A of the human speech cannot be substituted for the simulation. That is, the description of human speech is not and cannot become a simulation. A simulation is a different type of object from a description, the latter being a transformation of an observed object.

This is not to say that descriptions of real objects are not useful to the builder of simulations. Caution is necessary to understand the exact nature and purpose of the description and certainly it must not be assumed that the two can be substituted.

To return to the idea that researchers in speech synthesis and automatic speech recognition are in error about their assumptions as the nature of linguistics. I have now described two mistakes. The first is the assumption that the linguist's descriptive knowledge bases are algorithms for speech production or perception, and the second is the confusion between descriptive modelling and building simulations.

Both synthesis recognition systems profit from incorporating linguistic knowledge. But once again caution is necessary. In the work of so many researchers linguistic knowledge seems to mean *the knowledge linguists have*. It rarely means what it ought to mean: the knowledge base described by linguistics. But supposing we understand correctly what is meant by linguistic knowledge, how shall we incorporate it?

Fig.3

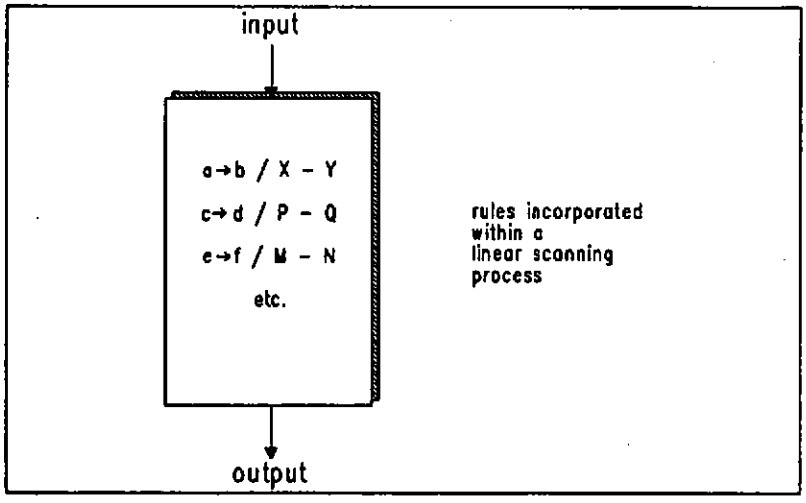


Fig.3 diagrams the phonological level of any of the current leading speech synthesis systems. A string from some higher level is input to a process. This process consists of a scan, often linear and therefore unprincipled, of a large set of rules taken directly from linguistics. Each

rule describes a transformation which must be applied to a particular phoneme in the input string if contextual conditions are satisfied. So for example, the string X*a*Y becomes X*b*Y: *a* becomes *b* if in the input string it occurs with a left context X and a right context Y, where X and Y can be null or strings of any length. In linguistics notation we might write *a* => *b* / X - Y. Often in such systems the entire set of rules has to be scanned for every element in every input string - a seemingly ridiculous procedure. Apart from such unprincipled and wasteful inelegance the procedure is not based on any sound theoretical consideration. The theory of human speech production would suggest a procedure accessing in a principled fashion a knowledge base of suitable rules, where the focus was on the method of access rather than on the knowledge base. I am not just playing with alternative layouts: the two approaches are *not* equivalent.

I shall not go into theoretical reasons for this assertion. But here is an example which makes the point. There are rules in the phonology and phonetics which describe the varying amounts of precision required in the articulation of speech sounds [5]. In the linguistic description these rules are labelled optional. Assume they are placed within the speech synthesis algorithm. Selection is made by scan of the rules for the item and its context: so all contradictory rules are selected. The inclusion of a meta-rule to the effect that in the case of optional rules only one may apply blocks all of the rules but one. Which? Is the choice to be random? Even a cursory examination of human speech reveals that the choice is not random but based on a reasoned decision.

Reasoned decision taking

Reasoned decision taking in human beings seems to rely on weighing up evidence or information from a number of sources. The evidence may constitute facts, which may shift in importance depending on circumstances, or beliefs. Reasoned decision taking seems to rest on the balance of probabilities surrounding these facts or beliefs. Use of the balance of probabilities at any one time is one of the mechanisms by which computation can take place when the evidence supplied comes from a large number of sources which may be different in type and when the evidence itself varies with respect to reliability.

It is reasoned decision taking based on evidence which is neither clear cut, nor guaranteed factual or stable and which relies on an assessment of probabilities which to a large extent distinguishes human behaviour from the usual form of machine behaviour. Simulation of reasoned decision taking falls within the domain of artificial intelligence. I am going to suggest that there is room in synthesis and recognition systems research for experimenting with a general artificial intelligence approach to some of the seemingly intractable problems we are coming across. Linguists who engage in simulation modelling rather than descriptive modelling are working within the area of artificial intelligence rather than pure linguistics.

At Essex University we have been experimenting with devices which can perform reasoned accessing of the knowledge bases described in linguistics. The knowledge bases are slightly different because they are intended for simulation rather than descriptive purposes. The kind of device which springs immediately to mind is the so-called expert system. Expert systems are designed to acquire evidence from their surroundings, to conduct a reasoning process and reach some conclusion by the selection of a particular goal from a number of given goals. Such a system for use at the phonological level in speech synthesis has been developed in our laboratory by Katherine Morton [6].

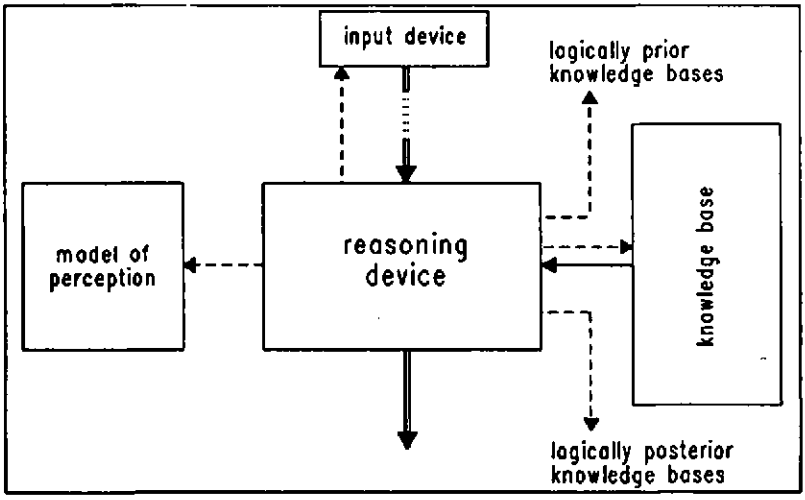
The goals of such a device may be a set of linked optional rules in the knowledge base. The task

is to select by reasoning not the correct rule to apply in the circumstances, but the *appropriate* rule. We are simulating an area of human behaviour where correctness is not the right term. By definition it could easily be the case that all the rules are correct (or they would not be in the knowledge base), but they are not equally appropriate on any one occasion. Each rule has assigned to it within the knowledge base an *a priori* probability weighting. That is, just as the knowledge base itself describes the native speaker's knowledge of the entire language and all utterances in that language, so these weightings indicate the probability of occurrence of each rule in the entire language. The reasoning device, or expert system, interrogates a number of sources of information. What sort of mood does the speaker (in this case the device determining what the system is to say) wish to convey? Does it believe the listener to be naïve in respect of the subject matter of the conversation? Does it believe the ambient noise or other factors in the environment merit special attention to precision in the utterance? Are there any reasons to suppose the listener will have any special difficulties in understanding the proposed utterance?

Other interrogation is made to sources of information within the synthesis system. Are there any special semantic considerations to be applied? Are the phonological units in the utterance, either separately or in combination, high or low in redundancy? Are there likely to be any special difficulties encountered by the lower-level phonetic component when attempting to execute the projected utterance? And so on.

Before it is answered each question has a predetermined effect on the *a priori* probabilities assigned to the options to be chosen among. Some of the questions will have only a binary possibility for their associated answers. Some questions will have a factual answer falling within a given range of possibility. Other questions will have answers informing of a degree of possibilitiy or probability. All the answers are computed with respect to their influence on the *a priori* probability weightings assigned within the knowledge base to the range of options. Finally one option will emerge with a resultant probability weighting greater than the others: and that is the one chosen as the most appropriate.

Fig.4



This is the principle behind our simulation of reasoned decision taking in speech synthesis.

Fig.4 illustrates simply what is going on. At the top of the figure there is the input device. It is here that a decision is taken as to what concepts are to be encoded as speech. An example of such a device would exist in an interactive database inquiry system such as the Alvey sponsored VODIS project. At the phonological level the expert system, as this level's reasoning device, accepts an input from the main concept-to-speech encoding algorithm. The task of the expert system is to transform this input depending on information accessed from the associated phonological knowledge base. That accessing is reasoned, as is selection from the knowledge base. The dashed links from the expert system indicate consultation or interrogation paths used in deciding what from other associated knowledge bases is relevant or appropriate for the job of transforming the main input to the main output.

There is one aspect of the system which has not yet been mentioned. Human beings as young children acquire the contents of their linguistic knowledge bases. In linguistic theory this is described as being accomplished by an iterative process performed by a language acquisition device producing successive grammars which progressively approach the mature grammar. Human beings as adults continually adjust their knowledge bases and indeed their strategies for decision taking. In other words they continually learn. The simulation must therefore have learning capability. Bridle [7] has discussed learning machines from the viewpoint of credible modelling of the mechanisms involved in learning. We are concerned with the functional counterpart of such mechanisms. So, although Bridle's machines are of a somewhat different type from what is discussed in this paper, we would be concerned with how the various weighting factors appearing in his model are arrived at in a principled way, and with the exact functional nature of the nodes or even labelling of nodes within the network.

Our proposed system is rather ambitiously a total one. We see no distinction between knowledge bases for speech synthesis or automatic speech recognition, nor any need for different types of mechanism to access them. The main synthesis and recognition algorithms may well be different, but we believe that better synthesis and recognition will emerge if, as in the human system, they are modelled as different modalities of the same overall device. Such a dual-mode device has many more possibilities internally for continuous updating of the weighting functions I have referred to above. So, for example, the device might ask itself: Was it the case that the utterance as I produced it evoked in the listener the desired or expected reaction? If the answer to this question is no, then some adjustment can be made automatically to some aspect of the decision taking processes within the device. In other words the system needs to have the means of detecting its own errors and in addition the means to repair the sources of those errors. In the field of artificial intelligence this kind of strategy is an aspect of what is known as knowledge engineering. That is, the acquisition and structuring of knowledge bases: in this case conducted automatically on a continuous basis.

Conclusion

This paper has discussed the nature of linguistic models and what they have to offer research in speech synthesis and automatic speech recognition. Linguistics provides a descriptive characterisation of the human knowledge base to support the encoding/decoding process of relating concepts with speech sounds, while saying nothing about the actual procedures involved. Speech synthesis and automatic speech recognition systems are simulations, not descriptions, focussing on the encoding and decoding algorithms. The direct substitution of sets of rules characterising a knowledge base for procedures is a mistake, as is the substitution of a description for a simulation. At the present time accessing of the knowledge bases in our simulations of speech production and perception is unreasoned and naïve. I have described an experimental method for reasoned access to the knowledge bases which is proving fruitful in producing a more natural and variable synthesised speech than currently available systems.

REFERENCES

[1] N. Chomsky, 'Formal properties of grammars', *Handbook of Mathematical Psychology* 2 (R.D. Luce, R. Bush and E. Galanter - eds.), New York: Wiley (1963)

[2] N. Chomsky, *Syntactic Structures*. The Hague: Mouton & Co. (1957)

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

[4] M.A.A. Tatham, 'An integrated knowledge base for speech synthesis and automatic speech recognition', *Journal of Phonetics* 13 (1985)

[5] M.A.A. Tatham and Katherine Morton, 'Precision', *Occasional Papers* 23, Essex University (1980)

[6] Katherine Morton, 'Intelligent speech synthesis', paper and demonstration given to the Leeds Experimental Phonetics Symposium, September 1986

[7] J. Bridle, paper presented to this conference.

Proceedings of The Institute of Acoustics

WORD-STRUCTURE REDUCTION RULES IN AUTOMATIC, CONTINUOUS SPEECH RECOGNITION.

Jonathan Harrington, John Laver and Doug Cutting

The Centre for Speech Technology Research, University of Edinburgh, Scotland.

1. INTRODUCTION.

A standard phonemic pronunciation dictionary such as Gimson's [1] will normally only include one phonemic entry for each lexical item. For the purposes of implementation in a feature-based, automatic speech recognition system[1], this is clearly inadequate since a given lexical item will often have many different possible pronunciations depending on factors such as tempo and context, as shown in Figure 1:

FIGURE 1: phonemicisations of the progressive modifications with increasing tempo to *audience*, *African* and *annual*.

	audience	*African*	*annual*	
Citation Form:	/oo d i @ n s/	/a f r i k @ n/	/a n y u(@) l/	slow, careful production
	/oo d y @̈ n s/	/a f r i k n/	/a n y u l/	
	/oo jh @ n s/	/a f r @ k n/	/a n y @̈ l/	
	/oo jh n s/	/a f r k n/	/a n y l/	
		/a f r k ng/		fast production

Using a Phonemic Rule Interpreter (described in [2]) encoded in INTERLISP-D on a Xerox 1108, the task has been to write a set of rules which derives automatically some of the fast speech forms of the lexical items in a 4,000 word lexicon implemented in a feature-based, connected speech recogniser. As in Figure 1, the rules should derive fast speech forms which are representative of the progressive modifications to lexical items with increasing tempo.

Initially, each lexical item has a single, phonemic entry, known as the *citation form*, and whenever a rule of fast speech applies, a *reduced form* is generated. In this paper, the criteria which were used in setting up citation forms and some of the rules which derive reduced forms from citation forms are discussed.

2. CITATION FORMS

In general, a citation form represents a slow and careful pronunciation of a given lexical item *in isolation*. Since 'careful pronunciation' is difficult to define, particularly across different speakers, two phonetic criteria were adhered to in setting up citation forms. First, a condition which applies to all rules that link citation and reduced forms is that such rules must be phonetically motivated. *Actually*, for example, has at least the following two possible pronunciations:

(2) /a k t y u l i/

(3) /a k ch u l i/.

But (2) would be chosen as the citation form since the derivation of /ch/ from /t y/ is phonetically motivated: /t/ is an alveolar stop, /y/ a palatal glide and /ch/, a palato-alveolar affricate, is the synthesis of the two. Second, apart from the 'linking /r/ rule' in some realisations in R.P. (Received Pronunciation) of the utterance *the idea is* as:

(4) /dh ii ai d i r i z/

fast speech rules do not insert phonemic segments. Insertion rules were therefore prohibited in the current rule set and this partly determined the choice of citation form. In the two possible productions of *anxious* in (5) and (6) for example, (5) would have to be chosen as the citation form and (6) would be a reduced form derived from (5) by rule:

(5) /a ng k sh @ s/

(6) /a ng sh @ s/

The examples discussed above show that the citation forms may represent productions which are in fact infrequently produced, or, at least, less frequently produced than some of their reduced forms. In the two phonemicisations of *actually* in (2) and (3), /a k t y u l i/ will probably occur less frequently than the reduced form /a k ch u l i/. For all citation forms, the condition was imposed that, although there was no restriction on their frequency of occurrence, they had to represent *possible* productions. Such a condition is a limit on the degree of 'abstraction' of citation forms and would exclude the underlying form /r i x t/ Chomsky & Halle [3] propose for *right* (since no R.P. speaker would ever produce this word as /r i x t/, where the /x/ is a voiceless velar fricative).

The inclusion of word boundaries (symbol #), syllable boundaries, primary stress and secondary stress in citation forms is necessary for the automatic derivation of reduced forms by rule. The rule which reduces /i/ in unstressed syllables to /@/ will

WORD-STRUCTURE REDUCTION RULES

apply in all environments unless /i/ occurs word finally: such a reduction rule must derive reduced form (8) from (7) but not (10) from (9):

(7)	/p r o b l i m/ (*problem*)	=	citation form
(8)	/p r o b l @ m/	=	reduced form
(9)	/s i t i/ (*city*)	=	citaton form
(10)	/s i t @/	=	illegal reduced form

A word-initial boundary symbol is required because the deletion of schwa before sonorants is more likely to occur when the schwa is word-medial. For example, schwa deletion is allowed to apply to (11) to derive (12):

(11)	/#d e t @ n ei sh @ n#/	(*detonation*)	=	citation form
(12)	/#d e t n ei sh n#/			

but, for the same tempo and degree of 'formality', should be prevented from applying on the citation forms (13) and (14):

(13)	/#d @ n au n s#/	(*denounce*)	=	citation form
(14)	/#t @ n ei sh @ s#/	(*tenacious*)	=	citation form

The inclusion of syllable boundaries is necessary to explain the tendency in fast speech for syllable *final* /t/ to glottalise, and possibly to delete, in (15) and (16):

(15)	/a t . l a n . t i k/	(*Atlantic*)	=	citation form
(16)	/a t . m @ s . f i @/	(*atmosphere*)	=	citation form

(the dot denotes a syllable boundary)

However, syllable *initial* /t/ will never delete in (17) to derive the reduced form (18):

WORD-STRUCTURE REDUCTION RULES

(17) /@.trakt/ (*attract*) = citation form

(18) /a.rakt/

Syllable boundaries are assigned on the basis of the 'maximum onset principle' [4]. In any $V_1C_nV_2$ sequence, where C_n stands for any number of consonants, C_n will be syllabified by V_2, provided that C_nV_2 is a phonotactically legal sequence for that language. If C_nV_2 is *illegal*, C_1 is syllabified with V_1 and the sequence $C_2C_3...C_n$ is considered as a candidate for syllabification with V_2, and so on. This algorithm results in a syllabification of *constrain* as:

(19) /k@n.strein/

The inclusion of primary stress in citation form entries (symbol '*') is necessary to prevent a rule which reduces /a/ to /@/ from applying on (20) to derive (21):

(20) /*a.t@m/ (*atom*) = citation form

(21) /*@.t@m/ = illegal reduced form

but to enable simultaneously /a/ in the first syllable of (22) to be reduced to /@/, thus deriving (23):

(22) /at.*lan.tik/ (*Atlantic*) = citation form

(23) /@t.*lan.tik/

i.e such a rule must reduce /a/ to /@/ only when /a/ appears in *unstressed* syllables. Secondary stress has been included for the same reason. If only primary stress were marked, the */a/ reduction rule* discussed above would apply to citation form (24) to derive the improbable reduced form (25):

(24) /a.lyu.*mi.ni.@m/ (*aluminium*) = citation form

(25) /@.lyu.*mi.ni.@m/

The rule can be blocked in this case by marking the first syllable of the citaton form for secondary stress. Both primary and secondary stress symbols, then, act as a filter to block reduction rules. In general, content words, but not function words, always include a syllable that is marked for primary, and optionally, secondary stress. Since the citation forms of *function* words are usually not marked for stress, is is possible to derive many appropriate reduced forms from their citation forms while simultaneously blocking the derivation of illegal reduced forms from citation forms of *content* words. In the function word *hadn't*, for example, four rules of *schwa deletion*, *coronal stop deletion*, */h/ deletion in unstressed syllables* and */a/ reduction in unstressed syllables* derive 15 reduced forms:

(26) /h a d @ n t/ (*hadn't*) = citation form
/a d @ n t/
/h a d @ n/
/a d @ n/
/h a d n t/
/a d n t/
/h a d n/
/a d n/
/h @ d @ n t/
/@ d @ n t/
/h @ d @ n/
/@ d @ n/
/h @ d n t/
/@ d n t/
/h @ d n/
/@ d n/

but the reduction of /a/ to /@/ in (20) and (24) is blocked, as described above.

3. REDUCTION RULES.

In writing reducton rules, it is necessary to set a limit on the degree to which citation forms are reduced. For example, in maximally fast speech, (27) has been reported [5] as a possible reduction of *San-Fransisco*:

(27) /s a m f s i s k ou/

It would be possible to write rules to generate such a reduced form from citation form (28); but an assumption has been made that a user of the connected speech recogniser is unlikely to produce speech with a tempo which would reduce the citation form to this extent[2]. On the other hand, since it is quite likely that such a user might produce the reduced form (29) at moderate to fast tempos:

(28) /s a n f r a n s i s k ou/ (*San-Fransisco*) = citation form

WORD-STRUCTURE REDUCTION RULES

(29) /s a m f r n s i s k ou/

a rule that assimilates /n/ to /m/ before /f/, a rule that reduces /i/ to /@/ in unstressed syllables and schwa deletion before sonorant consonants have been included in the current rule set for generating reduced forms.

The reduction rules which map citation forms onto the possible reduced forms can be classified into three different types: rules of *alternation*, *deletion* and *word-internal assimilation*; the latter can be further subdivided into *anticipatory* and *regressive* assimilation. The application of rules of *alternation* modifies the phonemic composition of a given lexical item but does not delete segments. Alternation rules relate, for example, /i/, /ii/ and /y/ in the three possible pronunciations of *associated* in (30) and /u@/, /uu/, /u/, /oo/ in four possible pronunciations of *curious* in (31):

(30) /@ s ou s ii ei t i d/ (*associated*) = citation form
/@ s ou s i ei t i d/
/@ s ou s y ei t i d/

(31) /k y u@ r i @ s/ (*curious*) = citation form
/k y uu r i @ s/
/k y u r i @ s/
/k y oo r i @ s/

Deletion rules, which delete one segment at a time, are only allowed to apply to /@/ in the case of vowels; therefore, the derivaton of the reduced form in (34) requires the prior existence of the reduced form (33) from citation form (32):

(32) /s i l i n d @/ (*cylinder*) = citation form

(33) /s i l @ n d @/

(34) /s i l n d @/

(32) - (34) embody a hierarchy from most careful pronunciation in (32) to the pronunciation produced at the fastest tempo in (34) as shown in Figure 1 above.

An example of word-internal *anticipatory assimilation* is the rule which relates the citation form (35) to the reduced form (36):

(35) /k r i s t m @ s/ (*Christmas*) = citation form

(36) /k r i s p m @ s/

An example of a rule of *regressive assimilation* is the assimilation of an alveolar nasal to a bilabial nasal when preceded by a labiodental fricative. This rule derives (38) from (37):

(37) /s e v n/ (*seven*)

(38) /s e v m/

The derivation of the maximum number of reduced forms from a given set of rules requires that the rules be ordered such that *alternation* rules precede *deletion* rules which precede *assimilation* rules. For example, the alternation rule relating citation form (39) to the reduced form (40):

(39) /k r i s t i @ n/ (*Christian*) = citation form

(40) /k r i s t y @ n/

(41) /k r i s t y n/

must apply before *schwa-deletion* which derives (41) from (40). This is because the schwa deletion rule is only allowed to apply if a *consonant* precedes the schwa (otherwise the impermissible form /i t a l i n/ would be derived from /i t a l i @ n/ (*Italian*), for example).

Assimilation rules must be ordererd after deletion rules; otherwise the rule of regressive assimilation that assimilates /n/ to /m/ in the context of a preceding labiodental could not apply to derive (44) (the derivation of (45) would also be blocked in this case since (45) is derived by a rule of anticipatory assimilation from (44)):

(42) /g uh v @ n/ (*govern*) = citation form

(43) /g uh v n/

(44) /g uh v m/

(45) /g uh b m/

Instead, *schwa-deletion* must first apply to (42) to derive (43); (43) then provides a context for the application of the regressive assimilation rule discussed above.

4. CONCLUSIONS.

The reduced forms described in this paper were generated from just over 30 complex rules that embodied approximately 450 single rules through the use of optionality and conjunctivity as described in [2]. The application of these rules resulted in the generation of 5300 reduced forms from the 4000 word citation form lexicon (i.e. the resulting lexicon consisted of 9300 citation and reduced forms). Currently, all citation and reduced forms are precompiled in a tree-structured lexicon in the connected speech recogniser. An alternative to precompilation is to include only citation forms in the tree-structured lexicon and to apply the reduction rules described in this paper in reverse on an incoming phoneme lattice prior to lexical access. Work is currently in progress to test the efficiency of this alternative implementation.

5. REFERENCES.

[1] A. Gimson, 'English Pronouncing Dictionary', 14th edition, originally compiled by Jones D. J.M. Dent & Sons Ltd.: London, (1984).

[2] D. Cutting and J.M. Harrington, 'Phongram: A Phonological Rule Interpreter', Proceedings of the Institute of Acoustics (This Volume), (1986).

[3] N. Chomsky and M. Halle, 'The Sound Pattern of English', Harper & Row: New York, (1968).

[4] D. Kahn, 'Syllable-Based Generalisations in English Phonology', Indiana University Linguistics Club: Bloomington, (1976).

[5] S. Kwasny, J. Dalby and R. Port, 'Rules for Automatic Mapping Between Fast and Slow Speech', Indiana University Computer Science Dept., Technical Report 175. (1985).

WORD-STRUCTURE REDUCTION RULES

6. NOTES.

1 The Machine Readable Phonemic Alphabet for R.P. used in this paper is shown below:

Phoneme (Vowels)		Phonemes (Diphthongs)		Phonemes (Consonants)	
/i/	*bid*	/ei/	*day*	/p/	*pea*
/ii/	*bead*	/ou/	*go*	/b/	*bee*
/e/	*bed*	/au/	*cow*	/t/	*tea*
/a/	*bad*	/ai/	*eye*	/d/	*dye*
/aa/	*bard*	/oi/	*boy*	/k/	*key*
/uh/	*bud*	/i@/	*beer*	/g/	*guy*
/@@/	*bird*	/e@/	*bare*	/m/	*me*
/@/	*the*	/u@/	*tour*	/n/	*name*
/o/	*pot*			/ng/	*sing*
/oo/	*port*			/f/	*fan*
/u/	*put*			/v/	*van*
/uu/	*boot*			/th/	*thin*
				/dh/	*then*
				/s/	*sea*
				/z/	*zoo*
				/sh/	*she*
				/zh/	*beige*
				/ch/	*chew*
				/jh/	*judge*
				/h/	*hat*
				/w/	*way*
				/y/	*yes*
				/l/	*lay*
				/r/	*ray*

2 Currently the subject of empirical investigation.