REACTION-TIME EXPERIMENTS IN THE STUDY OF SPEECH PROCESSING

D. B. FRY*

There is now a good deal of information available about the acoustic cues which are used in speech and their relation to the phonological system in certain languages. This knowledge is on the whole confined to the operation of single cues and to the initial stages of speech processing in the reception of speech. We understand in fact very little about the way in which acoustic cues for recognition are combined together and even less about the way in which the first stages of recognition are linked with the succeeding operations of linguistic processing.

The experiments described in this paper represent simply a first attempt to advance the enquiry by dealing with more complex situations and they are concerned more with finding a means of doing so than with producing far-reaching results.

First let us summarize very briefly the sequence of operations we may expect to take place in the reception of a spoken message: the acoustic input of speech is converted by the receptors into perceptual patterns which present a complex of features; these features are correlated with the acoustic cues and the listener has learned to make use of various combinations of cues in recognizing the sounds; the incoming sounds are assigned to the phonemic categories of the language on the basis of long-term *a priori* knowledge of the categories and short-term sequential information; the phonemic string forms the input for successive stages of further linguistic processing which yield morphemes and words which make up the message.

For the sake of convenience we can divide these operations into two main parts and refer to all the processing which depends directly on the acoustic cues and their combination as primary recognition, and the subsequent stages as linguistic processing. Either of these may call for operations of greater or less complexity. There will be contexts in which primary recognition is a simple operation, perhaps depending on the evaluation of a single acoustic cue; there will be others in which it is very much more complex, where it will be necessary for the listener not only to deal with a number of cues for a single phoneme recognition, but to process cues for a phoneme sequence where acoustic cues are interdependent and perhaps to operate at the same time upon cues for certain prosodic features. Similarly in linguistic processing, the

* University College London.

recognition of a phoneme string as a single-morpheme word may call for a comparatively simple piece of processing, whereas the string which forms a polymorphemic word, involving syntactic rules of a complicated kind, will require much more complex processing.

Although we at present know very little about the nature of these processing operations, it would be an error to assume that they are necessarily done serially in time. In fact, the most conspicuous feature of the functioning of the human brain is its very great capacity for doing many things at the same time; it seems able to employ the most intricate patterns of parallel working in such a way as to try out and discard many solutions to a problem in a very short time and hence to discover short cuts to correct solutions. Nonetheless, in the particular case of speech reception, it seems intuitively necessary that a complex piece of processing should take longer than a single simple operation. This is partly because the speech input is necessarily strung out in time, and a complex operation is likely to depend on information spread over a longer stretch of the acoustic continuum, and partly because of the hierarchical nature of language systems which requires that decisions on a lower linguistic level be made before processing on a higher level can be completed.

If it is the case that the more complex the speech processing, the longer it takes a listener to complete it, and if we could find some reliable means of determining the time that is required, we should have a way of gaining at least a qualitative idea of the complexity of the processing needed in a given case and perhaps eventually a criterion for distinguishing different processing operations. It is with this purpose in view that these reaction-time experiments have been begun. By setting a listener a variety of speech processing tasks and providing him with a means of signalling when he has completed the task, we may begin to gather evidence of the kind we have just referred to.

It is necessary first of all to establish whether the processing time does in fact appear to increase with the complexity of the task, and this can be done by beginning with very simple operations at the level of primary recognition, such as asking a listener to distinguish between words forming a minimal pair in his native language. The experiments reported in this paper do not in fact go much further than the exploration of this stage of the problem and the indication of some directions in which further progress seems likely.

THE TECHNIQUE OF REACTION-TIME EXPERIMENTS

At this point, it is necessary to say something briefly about the nature and the technique of reaction-time experiments. The essence of the method, as we have said, is to set the experimental subject some task, in this case a speech reception task, and to get him to signal, by pressing a button or a key, or perhaps by speaking, the moment when he has completed the task. The time interval between the arrival

of the speech stimulus at his ear and the making of the response is the reaction-time.

In these experiments, subjects responded by pressing a key. The method of asking the subject to speak back as a response was rejected because the object of the work was to investigate the operation of the speech reception mechanism, which we cannot suppose to be independent of the speech generating mechanism. To ask subjects to speak a response would be to set the speech generating mechanism working at the same time as the reception mechanism which is under observation. It seemed preferable to make the response a motor act unconnected with the speech mechanism and thus avoid the problem of determining from the data what part was played by the speech generating, as distinct from the speech reception, mechanism.

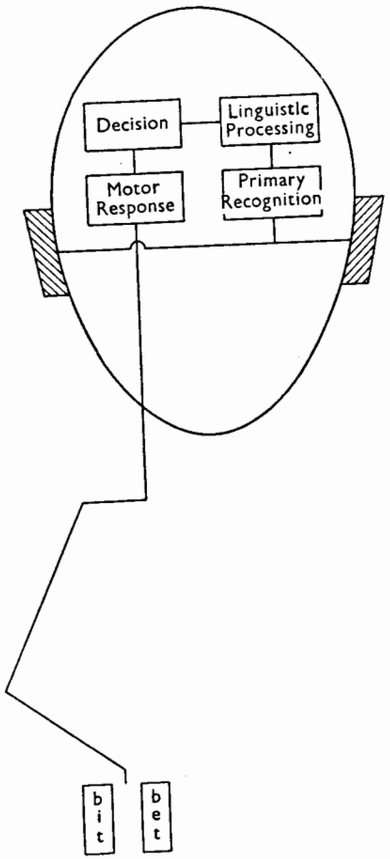

Fig. 1. Schematic diagram of response circuits involved in reaction-time experiments.

To take the simplest of the experiments as an example, the subject is asked to listen to words which are fed to him through telephone receivers; the words are in recorded natural speech, reproduced in good listening conditions so that they are easily recognized. A particular test, for example, consists of the words [*bit*] and [*bet*], occurring in random order; the subject has two keys, one marked *bit* and the other *bet* and his task is to press the appropriate key as soon as he has decided which word he has heard. This arrangement can be viewed schematically as it is shown in Fig. 1. The acoustic input to the two ears, suitably transformed by the hearing mechanism, provides the basis for the primary recognition of the sequence and this is followed by whatever linguistic processing may be necessary. This leads in turn to the decision as to which key is to be pressed and thus initiates the motor response, that is the neural command and the muscle action in the arm and finger which presses the key. When the speech input is changed, the operations represented in the left-hand side of the diagram remain essentially unchanged.

It must be noted, however, that reaction-times, in almost all circumstances, show considerable variability, even within one individual subject. If a subject is asked not even to make a choice but simply to press a key each time he hears a click in his

telephones, his response time (the simple reaction-time) will vary from moment to moment, from hour to hour and from day to day. That is to say that the operations represented by the left-hand side of the schematic diagram of Fig. 1 will make a contribution to the total reaction-time which is variable. In the present experiments the effect of this variability was minimised in the following way. In a given test-run, the subject was asked only to distinguish between two familiar words which were clearly audible to him, e.g. [*bit*] and [*bet*]. He was told to listen to each test item and as soon as he was certain which word had been spoken, he was to press the key marked with that word. He would then respond to a random sequence of the words [*bit*] and [*bet*] until he had made 50 responses to each word. This test-run takes about five minutes and whatever variations take place in the response mechanism in this short interval will tend to affect the responses to [*bit*] and [*bet*] equally, since they occur randomly during this period. Statistically significant differences in the reaction-times will consequently be due mainly to differences in the speech processing. This method of direct comparison was adopted throughout these experiments.

INDIVIDUAL DIFFERENCES IN SPEECH PROCESSING

A further source of variability, and for the purposes of these experiments a more important one, is to be found in the linguistic behaviour of individual listeners. Although the use of language for communication purposes depends upon the use of a common system, there is likely to be considerable individual variation certainly in the speed of processing and also to some degree in the sequence of operations. In decoding a spoken message, the listener's task is to arrive at the correct, that is the commonly accepted, solution; he retains a certain freedom as to the steps by which he reaches the solution and possible variations of this kind are important to our understanding of speech processing. It would therefore be a mistake to deal with the problem of individual variation by pooling reaction-times from a large number of subjects since by doing so we should lose some of the qualitative information that is of most interest to us. All the results given in this paper are therefore measurements from individual listeners who have made each a large number of responses to the test material.

REFERENCE POINTS FOR REACTION-TIME MEASUREMENTS

The first experiments to be described deal with very simple and basic considerations which are intuitively well understood. The technique is to ensure that the listener is in no doubt about the words he is going to hear, and to ask him to press the appropriate key as soon as he possibly can, but of course without making mistakes. If he is dealing with minimal pairs, the moment at which he can make a decision will clearly depend upon the point in the word at which the minimal

difference occurs. Transposing this into terms of primary recognition, at some point in the input the listener will pick out some acoustic cue or cues which will identify the member of the pair of words. Reaction-time might therefore be defined as the time interval between this moment and the subject's response. In practice we cannot be sure what cues the listener will use for the purpose nor at what moment they will appear in the input. For convenience, therefore, it is preferable to adopt some point of reference which is more easily defined and in all the measurements given here this point is the beginning of the word as determined from an oscillogram. This means the noise burst of a plosive sound, the beginning of friction noise in a fricative, the first voice cycle in the case of a voiced continuant, and so on.

If the reaction-time is then taken as the time interval between the beginning of the word and the pressing of the response key, we shall expect that this time will increase progressively as the minimal distinction between pairs of words occurs later and later in the words. This is in fact the case. If we take the word [*bit*] and place it in contrast first with the word [*pit*], then with [*bet*] and then with [*bid*], for each individual listener the reaction-time becomes progressively longer. This effect is presented in Fig. 2 where the spectrogram is that of the utterance [*bit*] used as the stimulus in the tests. The arrows indicate mean response times and all of them refer to the time taken by the subject to press the *bit* key, not to the time taken to respond to the contrasted words. Reaction-time is measured from the burst of the [*b*], and in these results, for subject N, the mean time taken to press the response key for [*bit*] when it is contrasted with [*pit*] is 325 msec., when contrasted with [*bet*], 359 msec. and when contrasted with [*bid*], 431 msec. Each value is the mean of 50 responses by the subject; the difference between the first two means is significant at a probability level less than 0.05, the difference between the second and third means is highly significant, at a probability level well below 0.001.

This effect, as we should expect, continues to appear when the sequence is extended to disyllables or trisyllables. Fig. 3 shows the mean reaction-times for the same subject responding to the contrast [*big*] — [*bid*], [*bigin*] — [*bigan*] and [*biginiŋ*] — [*biginə*]. The means for [*big*], [*bigin*] and [*biginiŋ*] are 430 msec., 522 msec. and 630 msec. respectively, differences between successive means being highly significant with a probability value well below 0.001.

These simple tests show one interesting fact, and that is that in the conditions of the experiment, subjects have no difficulty whatever in responding before a word or syllable is complete; the processing is capable of dealing with segments smaller than the whole word or syllable. This emerges more clearly if we normalise the results by referring the reaction-time in each case to the duration of the stimulus. Table 1 shows the means for subject D for the same set of contrasts, with reaction-time expressed as a proportion of the total duration of the stimulus. In only three cases [*big*], [*bid*] and [*biginə*], does the mean reaction-time exceed the total duration of the stimulus, and even here two of the values are almost equal to the duration.

Table 1. Subject D.

Mean Reaction-Time divided by Stimulus Duration

[*bit*]	0.666	contrasted with	[*pit*]	0.750
[*bit*]	0.748	.. ..	[*bet*]	0.734
[*bit*]	0.983	.. ..	[*bid*]	1.084
[*bid*]	0.930	.. ..	[*big*]	1.224
[*bigin*]	0.839	.. ..	[*bigan*]	0.748
[*biginiŋ*]	0.906	.. ..	[*biginə*]	1.030

Table 1. Mean reaction-times for an individual subject expressed as a proportion of the duration of the stimulus word.

PROCESSING OF CONSONANT CLUSTERS

We will now turn to tasks of a more complex kind, which still depend on the processing of acoustic cues and very little more. In these experiments subjects responded in the way already described to a whole series of contrasted words in successive tests. The list of these contrasts is given in Table 2, together with the mean reaction-times for subject C. It will be seen that the series builds up from minimal pairs with single initial consonants, to those with initial three-consonant clusters.

In the results that have been given so far we have been mainly concerned with differences in the reaction-time for the same word when it is placed in different contrasts. It is also possible that, when two words are contrasted, one of the two may produce reaction-times which are significantly shorter than those for the other. If this happens, it is in a general way a good indication that by this technique we are getting at differences in speech processing since in a single test other factors which might account for differences are balanced out.

Table 2 shows that in this series of tests there were six contrasts in which this subject gave significantly shorter reaction-times for one word than for the other. For example, when [*plei*] was contrasted with [*lei*], the mean was 346 msec. for [*plei*] and 394 msec. for [*lei*]. It was in fact generally true that the shortest reaction-times were provided by words beginning with a plosive when contrasted with words not beginning with a plosive. The sharp wave-front of the plosive onset is evidently a cue which is very rapidly processed. It is interesting, however, that in a contrast such as [*plei*] — [*lei*] the situation is asymmetrical in the sense that presence of the plosive wave-front is quickly registered yet its absence cannot be used as a cue; the reaction to [*lei*] has to wait for positive indications that the sound is [*l*]. The reaction-time for [*lei*] contrasted with [*plei*] is about equal to that for [*lei*] contrasted with [*rei*]. Since the principal acoustic cue for [*l*] and [*r*] is found in a rather slow frequency change in the second and third formants, it is not surprising that the reaction-times in these cases are rather long.

Table 2. Subject C.

	Mean of Reaction-times (msec.)			Mean of Reaction-times (msec.)
[*rei*]	417	contrasted with	[*lei*]	405
[*rei*]	387	.. .. ..	[*prei*]	370
[*rei*]	346	.. .. ..	[*sprei*]	372*
[*prei*]	332	.. .. ..	[*sprei*]	353***
[*prei*]	490	.. .. ..	[*plei*]	494
[*sprei*]	507	.. .. ..	[*strei*]	487
[*sprei*]	566	.. .. ..	[*splei*]	637***
[*lei*]	393	.. .. ..	[*plei*]	346***
[*lei*]	379	.. .. ..	[*slei*]	381
[*lei*]	365	.. .. ..	[*splei*]	385
[*plei*]	323	.. .. ..	[*slei*]	347***
[*plei*]	302	.. .. ..	[*splei*]	367***

* Significant difference between means within contrast, $p < 0.05$.

*** .. .. $p < 0.001$.

Table 2. A series of contrasts involving single initial consonants and initial consonant clusters with mean reaction-times for an individual subject.

Another interesting case is that of contrasts in which one word begins with friction noise and the other does not. In the case of [*plei*] versus [*splei*], we have for [*plei*] the shortest reaction-time in the whole series, 302 msec., which may well be the combined effect of the absence of friction and the presence of plosive onset. There is some tendency for the situation to be asymmetrical but with the absence of friction giving shorter times than its presence, although the differences do not always reach significance.

The most striking effect in this series of experiments, however, was seen in cases where consonant clusters of similar structure were contrasted with each other. In the examples we have just been discussing, we can see that the listener could arrive at a decision by giving his attention to a single acoustic cue—the sharp wave-front, the friction noise or the third formant frequency change. When two- and three-consonant clusters are contrasted with each other the situation seems to be much more complex. The contrasts [*prei*] — [*plei*], and [*sprei*] — [*strei*], [*sprei*] — [*splei*] produce a fairly dramatic increase in the time required for the task. Fig. 4 presents some of the relevant results in a somewhat clearer way. The spectrograms are analyses of two of the stimuli used in the tests, the left-hand one [*sprei*] and the right-hand,

[*splei*]. The arrows show on the time-scale the mean reaction-time for [*sprei*] and [*splei*] when they are contrasted with the word shown at each arrow. The mean reaction-time is 353 msec. when [*sprei*] is contrasted with [*prei*], and 372 msec. when it is contrasted with [*rei*]. We are here looking at means from different test-runs and there is no statistical significance in this difference. But when [*sprei*] is contrasted with [*strei*], the mean is 507 msec. and when it is contrasted with [*splei*], 556 msec. The difference between these values and the first two means is very highly significant, well beyond the level required for a probability of 0.001. The situation is paralleled when we look at the time taken to press the key for [*splei*] in different contrasts: in the [*plei*] and [*lei*] contrasts, the means for [*splei*] are 367 msec. and 385 msec., yet with [*sprei*] the mean is 637 msec.

The effect is equally marked in the case of the two clusters [*pr-*] and [*pl-*], shown in Fig. 5. The contrast [*prei*] — [*plei*] gives very long reaction-times, despite the fact that the other contrasts, with [*sprei*], [*rei*], [*splei*] and [*lei*], give some of the shortest reaction-times. Again the differences between the long and the short reaction-times are statistically highly significant.

It is very difficult to explain this effect except on the grounds that more complex processing is required in these cases. There is no doubt that the same acoustic cues are available in the various contrasts, and we can only conclude that the listener uses the minimum of cues that will enable him to make a decision in any given conditions. The processing is reduced to the simplest possible operations and when this means contrasting, for example, presence and absence of friction noise, the time required s very short. When the contrast is between [*pr-*] and [*pl-*], between [*spr-*], [*str-*] and [*spl-*], however, it seems that more acoustic cues have to be taken into account, the processing is more complex and the time required is correspondingly long. Although these clusters were all in initial position in the syllable, the mean reaction-times were almost invariably in excess of those needed for the contrast of final consonants in [*bit*] and [*bid*], and this despite the fact that the duration of the stimulus word [*prei*] was identical with that of the word [*bit*].

PROCESSING OF MORPHEME BOUNDARIES

The experiments already described have all been confined to the processing of acoustic cues, with no demand for subsequent linguistic processing. It would clearly be very valuable if the reaction-time technique could be used to explore higher levels of processing. This last section deals with one set of preliminary expreiments carried out to see whether there are possibilities in this direction.

These tests involved three contrasts in which a morpheme boundary occurred. [*pik*] — [*piks*], [*pik*] — [*pikt*], [*piks*] — [*pikt*]. Several factors make it difficult to draw firm conclusions from these tests, apart from the fact that many more contrasts need to be investigated. In the nature of the English language system, such morpheme boundaries involve word-final consonants or consonant clusters so that

the acoustic cues on which the processing is based come towards the end of the word and reaction-times measured from the beginning of the word will in any case appear long. Furthermore, it is difficult if not impossible to find in English minimally contrasted sequences of which one involves a morpheme boundary and the other does not without having recourse to orthographic differences which are doubtfully represented at the phonetic level.

A typical set of mean reaction-times for the [*pik*] contrasts is given in Fig. 6. It is true that all the mean reaction-times are greater than the duration of the stimulus word, sometimes by a considerable amount, but this is not unexpected in view of the position of the contrasting sounds. One might guess that if listeners were operating on the very simple acoustic basis which was noted in the case of [*prei*] — [*sprei*], etc., the response might well come sooner than it does here. On the evidence of this one set of contrasts, however, it is impossible to say whether the presence of the morpheme boundary is affecting the processing time or not.

This question along with many others, including the whole field of the processing of prosodic features, may be answered by further experiments along these lines. The preliminary results reported in this paper do at least suggest that there is much to be learned about speech processing by measuring how long it takes listeners to carry out a particular set of operations.

DISCUSSION

Hill:

I do not wish to question, but to say why I feel this one of the most significant phonetic papers I have heard in years. We have been hearing lately a great number of statements that processing of what we hear may not be linear, but all at once even reversed. This is the first time I have heard anyone deal with the problem directly.

Lehiste:

Was there any learning effect observed in the responses?

Newell:

Would not the fact that you were using only one acoustic waveform for each stimulus, rather than different attempts at the same words by the same speaker, decrease the reaction time, due to recognition of slight differences in the phonetically similar parts of the stimulus.

Fry:

Ad Lehiste: I have not examined systematically the statistical variation in reaction-time within each test, but from mere inspection it seems that if there is any degree of learning in making the responses it takes place very rapidly indeed. The subject received no warning signal when the first stimulus in a test was about to arrive, and for this reason the very first reaction-time was sometimes quite long, but apart from this, it was never possible to note by inspection that the first few times were longer than succeeding ones. Furthermore, there were two test runs for

each minimal pair and there was no tendency for the mean of the first 25 responses to a given word to be greater than the mean of the second 25. This did sometimes happen but about equally frequently the second was greater than the first. Had there been a pronounced learning effect lasting over some minutes, there would have been a systematic trend towards shorter reaction-times in the second of each pair of test runs.

Ad Newell: While I think it is possible that if the amount of variation in the stimulus words were increased, the reaction-time might increase. I do not think that this is beyond question. If by this reaction-time method we are really able to get a measure of the acoustic-linguistic processing time, then it is conceivable that the introduction of a certain degree of variation in the stimulus may not lead to a significant increase in reaction-time.

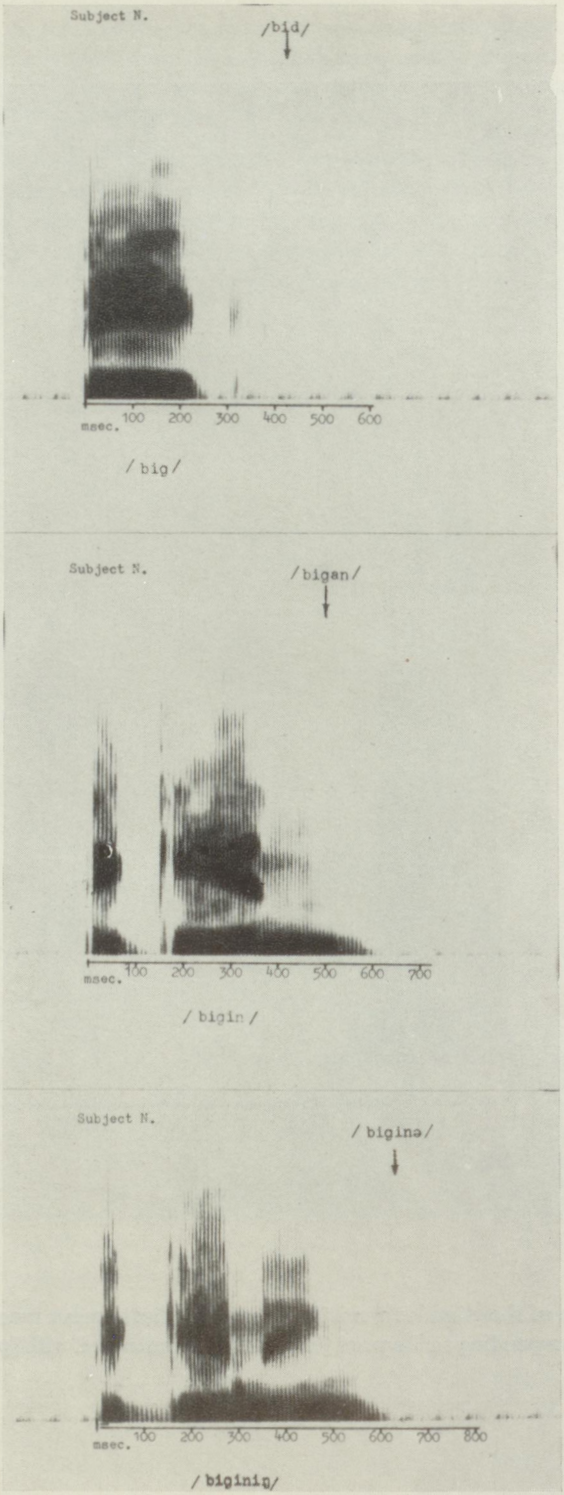

Fig. 3. Spectrograms of the stimulus words [*big*], [*bigin*] and [*biginiŋ*]. Arrows indicate mean reaction-times for an individual subject responding to the words [*big*], [*bigin*] and [*biginiŋ*] when contrasted with [*bid*], [*bigan*] and [*biginə*] respectively.

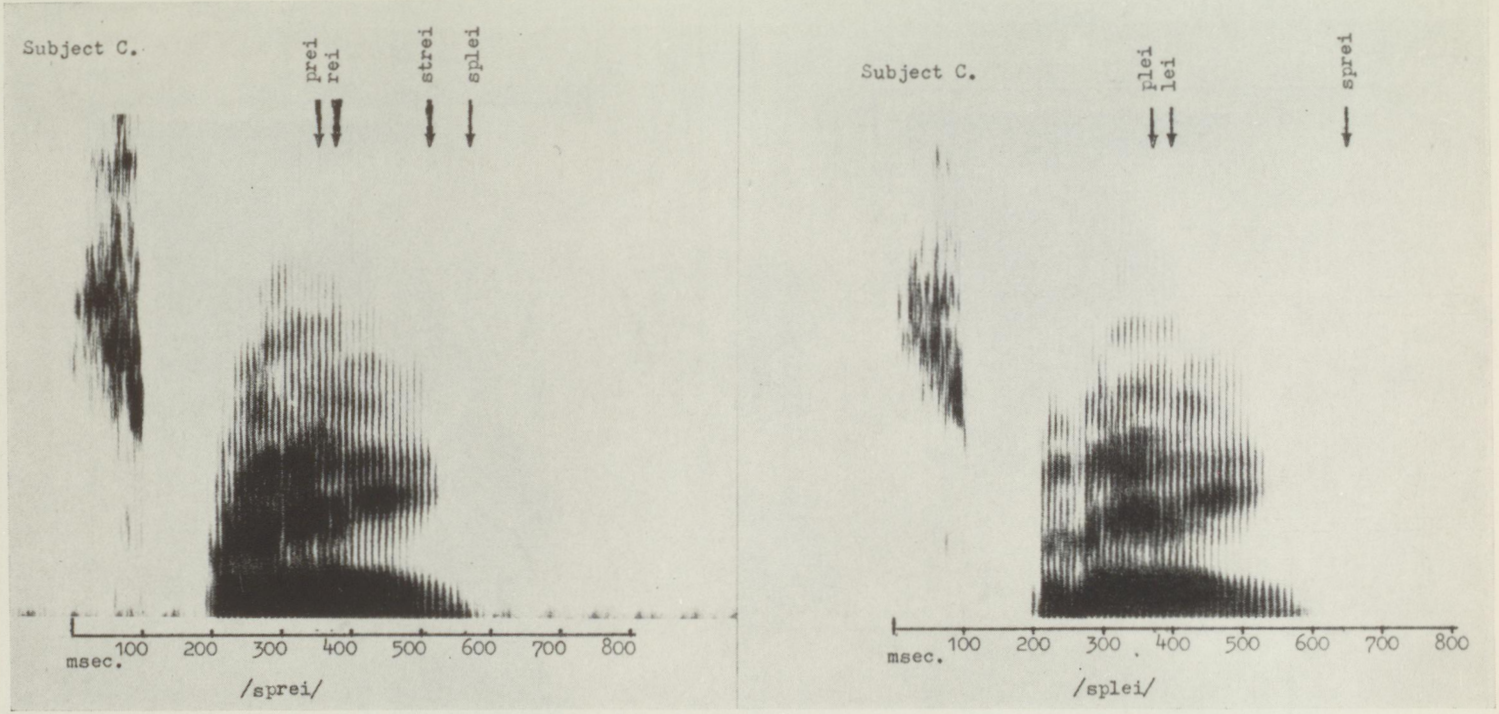

Fig. 4. Spectrograms of the stimulus words [*sprei*] and [*splei*]. Arrows indicate mean reaction-times for an individual subject responding to these two word in the contrasts [*sprei*] — [*prei*], [*sprei*] — [*rei*], [*sprei*] — [*strei*], [*sprei*] — [*splei*], and [*splei*] — [*plei*], [*sprei*] — [*lei*] and [*splei*] — [*sprei*].

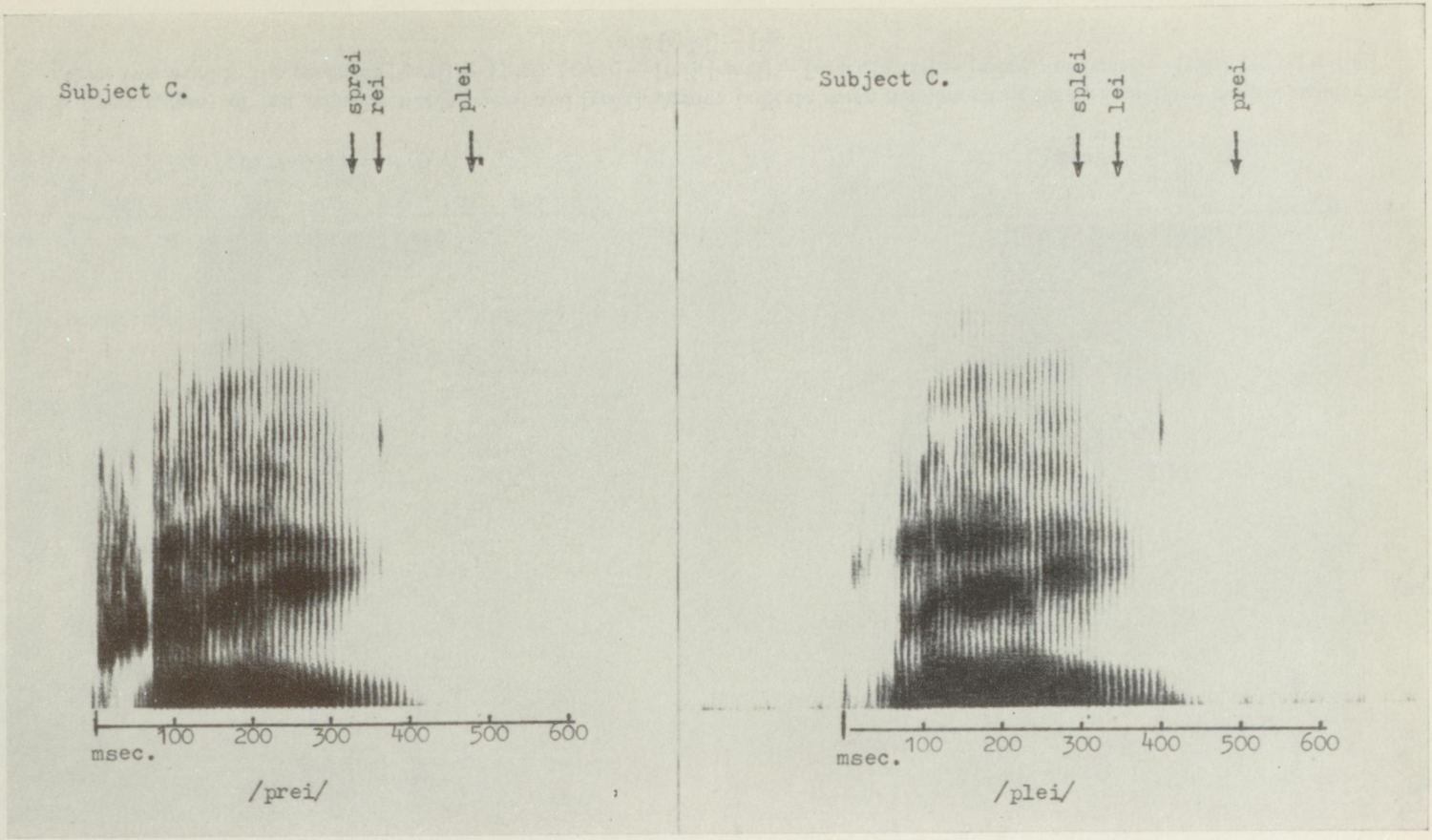

Fig. 5. Spectrograms of the stimulus words [*prei*] and [*plei*]. Arrows indicate mean reaction-times for an individual subject responding to these two words in the contrasts [*prei*] — [*sprei*], [*prei*] — [*rei*], [*prei*] — [*plei*], and [*plei*] — [*splei*], [*plei*] — [*lei*] and [*plei*] — [*prei*].

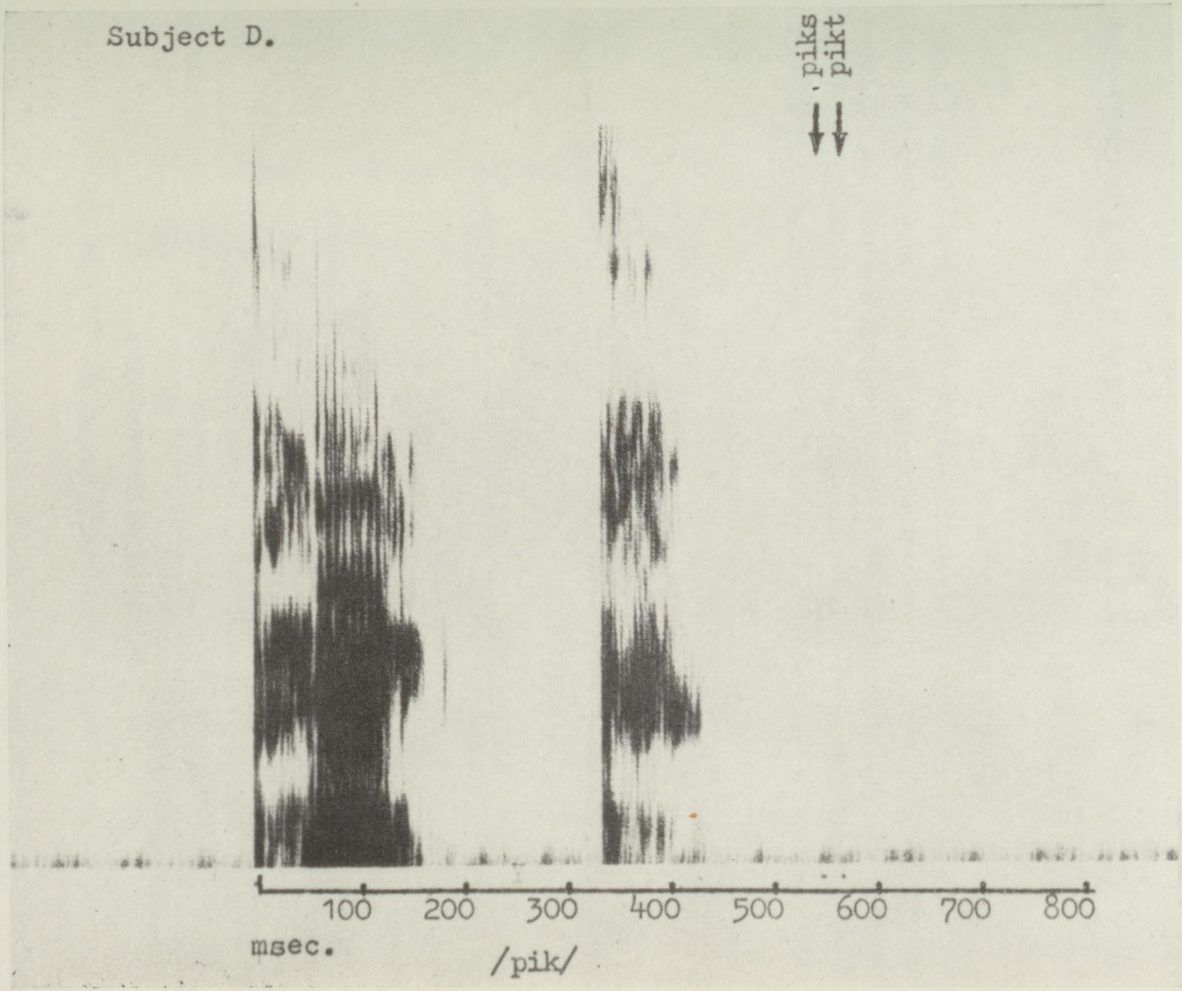

Fig. 6. Spectrogram of the stimulus word [*pik*]. Arrows indicate mean reaction-times for an individual subject responding to this word contrasted with the verb forms [*piks*] and [*pikt*].