Computer Speech and Language (1993) 7, 211–228

Phoneme detection as a tool for comparing perception of natural and synthetic speech

Andrew J. Nix, Gita Mehta, Julie Dye and Anne Cutler*

*Cambridge University Engineering Department; *MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF*

Abstract

On simple intelligibility measures, high-quality synthesiser output now scores almost as well as natural speech. Nevertheless, it is widely agreed that perception of synthetic speech is a harder task for listeners than perception of natural speech; in particular, it has been hypothesized that listeners have difficulty identifying phonemes in synthetic speech. If so, a simple measure of the speed with which a phoneme can be identified should prove a useful tool for comparing perception of synthetic and natural speech. The phoneme detection task was here used in three experiments comparing perception of natural and synthetic speech. In the first, response times to synthetic and natural targets were not significantly different, but in the second and third experiments response times to synthetic targets were significantly slower than to natural targets. A speed-accuracy tradeoff in the third experiment suggests that an important factor in this task is the response criterion adopted by subjects. It is concluded that the phoneme detection task is a useful tool for investigating phonetic processing of synthetic speech input, but subjects must be encouraged to adopt a response criterion which emphasizes rapid responding. When this is the case, significantly longer response times for synthetic targets can indicate a processing disadvantage for synthetic speech at an early level of phonetic analysis.

1. Introduction

The primary aim of speech synthesis is to produce speech which can be used for communicative purposes—that is, speech which is intelligible. Perceptual measures of synthetic speech, and comparisons between systems, have therefore concentrated on measuring intelligibility. Although the intelligibility of some systems remains low in comparison to natural speech (Hoover, Reichle, Van Tasell & Cole, 1987), the best systems produce speech which is highly intelligible and, in fact, not very much less intelligible than natural speech. Logan, Greene & Pisoni (1989), for example, tested the output from ten synthesizers, the best of which produced speech which was 96·75% intelligible; this was significantly worse than the 99·5% intelligibility of natural speech in their study, but it is far closer to the natural speech result than to the 64·5% and 72·5% intelligibility produced by the worst performers among the synthesizers they tested.

* Correspondence to Dr A. Cutler

0885–2308/93/030211 + 18 $08.00/0

It might seem reasonable to conclude that because good-quality synthesized speech is highly intelligible it should be as easy for listeners to process as natural speech. But in fact, listening to synthetic speech appears to place greater capacity demands on the listener's processing resources than listening to natural speech; for instance, concurrent recall tasks have a larger adverse effect on processing of synthetic speech than on processing of natural speech (Luce, Feustel & Pisoni, 1983; Lee & Nusbaum, 1989). The amount of information which is retained from presented text (as measured by accuracy of sentence verification, or correct responses to comprehension questions) tends to be significantly reduced if the text is presented in synthetic as opposed to natural speech (for a review see Ralston, Pisoni & Mullenix, in press).

There are many ways in which synthetic speech does not sound like natural speech. One noticeable aspect is speech prosody, which is generally much less varied in synthetic than in natural speech. Certainly improvement in prosodic information leads to synthesis output which is preferred by listeners (Silverman, 1987; Terken & Lemeer, 1988), and which appears to be processed more efficiently (Larkey & Danly, 1983). However, it has been argued (Pisoni, Nusbaum & Greene, 1985) that the added difficulty of listening to synthetic speech does not occur at larger, prosodic, levels but at the lower level of cues to phonetic identity; synthetic speech places greater demands on the earliest stages of perceptual processing. In natural speech phonetic information is often redundantly specified; in synthetic speech the cues are sparser, which implies harder work for the phonetic processor. This perceptual disadvantage is then assumed to carry through all levels of language processing.

In support of this hypothesis, Nusbaum, Dedina and Pisoni (1984) demonstrated that the pattern of perceptual confusions for consonants in CV syllables differed in natural and in synthetic (even high-quality synthetic) speech. The processing capacity demands made by listening to synthetic speech have also been specifically explained in these terms (Waterworth & Holmes, 1986; Lee & Nusbaum, 1989). Pisoni (1981) likewise invoked phonetic processing difficulty as the explanation of the disadvantage for synthesized vs. naturally spoken words which he observed in auditory lexical decision and naming. In his first study, listeners classified auditorily presented strings as words or nonwords; response time was significantly slower for the synthesized strings. However, real words were responded to faster than nonwords to the same degree in both natural and synthetic speech. The same pattern of responses occurred in his second study using the naming task (in which listeners have to repeat the presented words): responses were faster for natural speech than for synthetic, and for real words than for non-words, but the effects did not interact. Pisoni argued that the identical effects of lexical status were evidence that the type of processing being carried out on natural and synthetic speech was the same; this allowed him to infer indirectly that the listener's difficulty is therefore probably located in the early stages of perceptual analysis of the speech signal.

Pisoni's study is one of only few in which response time (RT) measures have been applied to the processing of synthetic speech. RT measures are assumed to reflect fluctuations in processing difficulty more directly than any measure taken after processing is complete (Levelt, 1978). Therefore they offer a logical next step in perceptual evaluation of synthetic speech. As intelligibility scores for high-quality speech approach those for natural speech, it is appropriate to turn to measures which look more closely at speech processing as it occurs. This argument has been made strongly by Pisoni, Manous and Dedina (1987) and by Ralston, Pisoni and Mullenix (in press). Ralston *et al.* (in press) summarize a number of studies in which responses were faster for natural than for

synthetic speech stimuli in such tasks as sentence comprehension, monitoring for target words, and sentence verification. Pisoni, Manous and Dedina also found that sentence verification responses were faster for natural than for synthetic speech input, even in sentences controlled for relative intelligibility. Lee and Nusbaum (1989) and Ralston, Pisoni, Lively, Greene and Mullenix (1991) confirm that word monitoring RTs are slower for synthetic than for natural speech.

None of the above RT measures directly assess the difficulty of phonetic processing; but measures of this do exist. Detection of target phonemes—"phoneme-monitoring"—is a task developed by Foss (1969), in which listeners hear a sentence or list of words and press a response key when they detect a word beginning with a specified target phoneme. The nature and applications of this task have been extensively explored (see, e.g., Cutler & Norris, 1979; Foss & Gernsbacher, 1983; Cutler, Mehler, Norris & Segui, 1987); but it has not been used to assess the difficulty of processing synthetic vs. natural speech. It seems reasonable to expect, however, that if phonetic processing of synthetic speech causes listeners difficulty, then this difficulty will produce a performance decrement in the phoneme detection task. Phoneme detection is a simple experimental task and would lend itself well to inclusion in the armoury of speech technology assessment methods. Moreover, it would provide a more direct reflection of phonetic processing difficulty than the indirect arguments from higher-level measures made, for instance, by Pisoni (1981). In the present study, therefore, we tested the applicability of the phoneme detection task to the assessment of perception of natural vs. synthetic speech. In Experiment 1 we used the highest quality synthetic speech we could find; this was in fact DECtalk, the speech system which scored 96·75% in Logan *et al.*'s (1989) comparative tests.

2. Experiment 1

2.1. Method

2.1.1. Materials

An hour and a quarter of spontaneous, unprompted conversation was recorded in an anechoic, sound-dampened room between two male speakers of British English. A transcript of the recording was made, and 53 utterances, all by the same speaker, were chosen for use in the experiment: 30 for use as experimental sentences, the remainder as fillers, warm-up and practice sentences. The experimental sentences are listed in the appendix. Each of the stop consonants /p/, /t/, /k/, /b/, /d/ and /g/ was target in five experimental sentences. In any experimental sentence, the target phoneme occurred only once in word-initial position. In fillers (intended to ensure that subjects did not just wait for the end of each sentence and respond then) the designated target phoneme did *not* occur word-initially anywhere in the sentence.

A week after the conversation was recorded, the speaker whose utterances had been selected returned to the laboratory and recorded from a written text all 53 isolated sentences, in randomized order. The sentences were read with natural emphasis and intonation, and were recorded in the same sound-dampened room, and with the same recording equipment, as the spontaneous recording. The speaker did not know which sentences contained targets, or what the targets were.

The 53 sentences were then synthesized with DECtalk (Digital Equipment Corporation, 1983), using the male voice output option "Paul".

Three experimental tapes were constructed, each beginning with the five practice sentences in all three modes (uttered spontaneously, read naturally, and synthesized). Each tape then contained three blocks of 16 sentences, each block consisting of four warm-up sentences (responses to which were not recorded), ten experimental sentences, and two fillers which were placed randomly among the experimental sentences. Each tape contained only one occurrence of each experimental sentence, and all three speech modes. On Tape 1 Block A (sentences 1–10 in the appendix) was spoken spontaneously, Block B (sentences 11–20) was read, and Block C (sentences 21–30) was synthesized; on Tape 2 Block A was read, Block B was synthesized and Block C was spoken spontaneously; on Tape 3 Block A was synthesized, Block B spontaneous and Block C read. Thus the three types of speech materials used in this study were fully counterbalanced. Note, however, that one type of material—spontaneous speech—will not further be considered here. The results of the comparison between spontaneous speech and read speech have previously been published (Mehta & Cutler, 1988). The comparison of interest here is that between synthetic and natural speech; the form of natural speech to be compared with synthetic is (as in other such studies) the naturally read speech.

The REVOX tape recorder's built-in slide synchroniser was used to record a brief tone aligned with the release burst of each word-initial target phoneme, and also with the end of each sentence. The tone was inaudible on the speech channel which the subjects heard; the slide synchroniser head, positioned between the two channels of the tape, both recorded and detected the tone. All experimental sentences were digitised and the difference between the position of the timing tone and the actual onset of the burst of each target phoneme was measured to the nearest msec. RTs were then adjusted for this difference to give times from phoneme onset.

Because the experimental sentences had been spontaneously spoken, they were not systematically controlled. However, it seemed desirable to adopt the approach of Pisoni (1981) and examine some effects known to occur with this task which might offer the possibility of detecting any qualitative difference in listeners' processing of the two speech modes. The experimental sentences offered sufficient variation to enable post-hoc analyses of five effects previously reported in the phoneme detection literature. The variables are not orthogonally distributed in the materials, but this is true of both speech modes; effects can be analysed in terms of their *relative* strength in each speech mode. The five variables were:

(1) *Transition probability.* Morton and Long (1976), Dell and Newman (1980) and Mehler and Segui (1987) reported that phoneme targets on contextually predictable words are detected faster than targets on unpredictable words. Following Morton (1967), we measured transition probability by presenting each experimental sentence up to, but not including, the target bearing word to 20 native speakers (not subjects in the experiments), who were asked to continue each sentence with the first word(s) that occurred to them. For instance, in ten replies "Have you got a very big record ..." produced "collection"; so as "collection" was indeed the target bearing word, that sentence was scored ten. Scores above 5 (i.e. 25% response) were deemed high on transition probability, below 5 low.

(2) *Preceding word length.* Mehler, Segui & Carey (1978) and Newman & Dell (1978) found that target phonemes preceded by longer words were detected more rapidly than targets preceded by shorter words. We compared RTs to targets preceded by monosyllables vs. longer words (in these materials the range of the latter was two to five syllables).

(3) *Position of the target bearing word.* RTs tend to be faster the later the target-

bearing word occurs in a sentence (Foss, 1969; Shields, McHugh & Martin, 1974; Cutler & Fodor, 1979). We contrasted targets occurring in syllable positions 1 to 5 (early) with targets occurring in syllable positions 6 to 16 (late).

(4) *Sentence accent.* RTs are faster on accented than on unaccented words (Cutler, 1976; Cutler & Foss, 1977; Shields, McHugh & Martin, 1974). Four listeners (the second and fourth authors and two colleagues) transcribed the accent patterns of the experimental sentences. We contrasted sentences in which the target-bearing word was judged to be accented by two or more listeners with sentences where the target-bearing word was judged accented by no or only one listener.

(5) *Syllable stress.* Taft (1984) found that initial phonemes were detected more rapidly if the syllable in which they occurred was stressed; post-hoc examination of the materials used by Cutler and Foss (1977) revealed the same result. In seven of the 30 present sentences the target occurred in a word with a weak initial syllable (e.g. *collection, believe*); in the remaining 23, the target occurred in a strong syllable (e.g. *difficult, power*).

2.1.2. Subjects

Thirty-nine subjects were tested, all but one from the Medical Research Council's Applied Psychology Unit subject panel. The subjects were between 22 and 45 years of age and were native speakers of standard British English. The results of three subjects were discarded because they missed one-third or more targets in any one block or because of an unacceptably high error rate in the recognition test. Of the remaining 36 subjects, twelve heard each tape, and of these twelve, four heard each of three speech mode orders. All subjects from the APU subject panel were paid for their participation in the experiment, which lasted approximately 20 min.

2.1.3. Procedure

Subjects were tested individually in a sound-dampened room. They were given written instructions which emphasized speed of response, but also accuracy in responding only to targets in word-initial position. They were also instructed to pay attention to the meaning of the sentences as they would receive a recognition test later.

Subjects heard the speech materials over headphones. The target for each sentence appeared on a VDU screen in front of the subject. Target presentation, timing and response collection were controlled by a DEC PDP 11/23 minicomputer running TSCOP (Norris, 1984). The timing tone aligned with the onset of each target phoneme was detected by the tape recorder's slide synchroniser, which triggered a timer in the computer; the timer was stopped by the subject pressing a response key.

Subjects first heard the practice set, in the speech mode corresponding to the first block of experimental sentences which they would receive. The three experimental blocks followed, with short breaks between blocks. To control for effects of order of presentation of speech modes, the three blocks within each tape were presented in three counterbalanced orders, one order being heard by one-third of the subjects who heard each tape.

A recognition test was given to each subject after the experiment. Half the subjects received a test of 20 sentences, ten of which they had heard exactly, while the other ten were constructed by putting together phrases from more than one (experimental or filler) sentence. This is the standard form of recognition test used in phoneme detection experiments, and is intended to provide a rough check that subjects are indeed attending to the content of the speech material. The test administered to the remaining subjects

consisted entirely of sentences actually heard, selected roughly equally from the three subsets; thus about a third of them had been heard by any given subject in each of the three speech modes. Subjects who received this test were told specifically that they had heard only some of the sentences in the test and not others. In both instances, subjects were required to respond "Yes" or "No" to whether or not they had heard the precise wording of the sentence in the test. This second test was intended to ascertain whether there was any difference in recall as a function of speech mode.

2.2. Results and discussion

The mean score on the recognition test for the first 18 subjects was 60%. Sentences which subjects had heard in their synthesized form were somewhat less well recalled (43·1% error) than sentences heard as natural speech (37% error): however, this difference was not significant ($t(17) = 1{\cdot}18$, $P > 0{\cdot}1$). The other form of recognition test, received by the remaining 18 subjects, also failed to show statistically significant effects of speech mode on recall. In this case, the overall percent correct was 62%, and subjects made 7% more errors on sentences which they had heard in synthesized form than on sentences which they had heard as natural speech.

On average subjects missed 0·97 targets in natural read speech and 1·22 targets in synthetic speech; the difference was also not statistically significant.

RTs underwent separate analyses of variance with subjects (*F1*) and sentences (*F2*) as random factors. The overall mean RT to targets in synthetic speech was 579 msec, to the same targets in natural speech 558 msec. The 21 msec difference was not significant on either analysis. There was also no significant effect of the order in which subjects heard the speech modes, nor did this effect interact either with the speech mode factor or with which tape (i.e., which combination of sentence blocks with speech modes) the subjects had heard.

Thus this experiment found no overall effect of speech mode: phoneme detection RT was not significantly different to targets in synthetic vs. naturally produced speech. Furthermore, post-hoc examination of the five previously reported effects also showed little difference between the two speech modes. The mean RT to targets on high transition probability words was faster than to targets on low transition probability words by 33 msec for natural speech and by 23 msec for synthetic speech; in neither case, however, did the effect reach statistical significance. Targets preceded by longer words were responded to faster than targets preceded by monosyllables; however, the difference again did not reach statistical significance. The difference was much larger for natural speech (46 msec) than for synthetic speech (4 msec); the interaction of speech mode with preceding word length was significant in the analysis by subjects ($F1[1,27] = 5{\cdot}47$, $P < 0{\cdot}03$), but not in the analysis by items ($F2 = 1{\cdot}45$). Neither sentence accent nor syllable stress showed any significant effects with either speech type. The only significant effect observed was for position in the sentence: late targets were detected significantly more rapidly than early targets, both for natural speech (by 123 msec) and for synthetic speech (by 96 msec); this effect was significant in both analyses ($F1[1,28] = 13{\cdot}97$, $P < 0{\cdot}001$; $F2[1,28] = 13{\cdot}65$, $P < 0{\cdot}001$). The effect was equally significant in each speech mode ($t[35], = 3{\cdot}6$, $P < 0{\cdot}001$ for natural speech, $t[35] = 3{\cdot}24$, $P < 0{\cdot}003$ for synthetic speech).

Somewhat surprisingly, then, this experiment showed no significant difference between RTs to phonemes in synthetic as opposed to naturally produced speech. The

overall mean RTs were somewhat slower for synthetic speech but not significantly so, and the pattern of results on the five independent variables which we analysed was almost identical for the two speech modes. This latter result agrees with previous findings, by, for example, Pisoni (1981) and Pisoni, Manous and Dedina (1987) that the way in which listeners process speech does not differ as a function of whether it is presented in natural or synthesized form; but the failure to find a significant overall RT disadvantage for synthetic speech is in disagreement with all the RT studies cited above.

However, one noticeable aspect of the results is that the RTs are long by comparison with other phoneme detection studies using sentence materials (mean RT in Cutler & Foss's [1977] study, by comparison, was 425 msec, and Cutler & Fodor [1979] 381 msec). A possible explanation is that subjects may have found it hard to comprehend the materials out of their original context, and this difficulty may have masked differences in the ease with which they processed natural vs. synthetic speech. Accordingly, in Experiment 2 we used materials which were constructed to stand alone as isolated sentences and did not require contextual reference for their interpretation. By constructing our materials to prior specification we could also manipulate relevant characteristics of the sentences in a more systematic way than in Experiment 1. We chose to manipulate the transition probability, position in sentence, and syllable stress pattern of the target-bearing words.

3. Experiment 2

3.1. Method

3.1.1. Materials and procedure

Twenty-four target words were chosen, each beginning with one of the five stop consonants /b, d, g, p, k/. The words formed pairs; both members of a pair began with the same phoneme and had the same stress pattern. Half of the pairs began with strong syllables (e.g. *garlic/garbage*) and half with weak syllables (e.g. *bananas/belongings*). The members of a pair were matched for number of syllables and frequency of occurrence (Kučera & Francis, 1967). Each word was embedded in a relatively plausible sentence context. For half of the pairs, the critical word occurred early in the sentence (preceded by no more than five syllables) and for the other half it occurred late (preceded by at least ten syllables). Low transition probability sentences were then constructed by exchanging each critical word with its pair. The experimental sentences are listed in the appendix.

A pre-test confirmed that the two sets of sentences differed on mean transition probability. As for Experiment 1, subjects who did not take part in the experiments provided sentence completions. 18 subjects completed sentence frames which contained the context preceding the target-bearing word, while another 17 completed frames which contained this context plus the initial letter of the target item. The latter procedure produced more than twice as many completions which were actually the target word than the former procedure did. However, both procedures produced a significant number of correct completions for high transition probability sentences but virtually none for the low transition probability versions.

24 filler and warm up sentences, some with no occurrence of the specified target, and eight practice sentences were also constructed. All sentences were recorded by a male native speaker of standard American English (selected to have voice quality resembling

that of the American-accented synthesizer output). The circumstances of the recording were as for Experiment 1.

All sentences were synthesized using the MITalk text-to-speech system (Allen, Hunnicutt & Klatt, 1987). MITalk is in most respects identical to DECtalk, so that the synthetic speech of Experiment 2 should be comparable to that of Experiment 1; MITalk was chosen here because Experiment 2 again contained a third speech mode, which could only be generated via MITalk. This third type was synthetic speech in which phoneme durations had been manipulated by hand. Once again, however, as the focus of this paper is on the comparison between natural read speech and high-quality synthesizer output, this third speech mode will not be discussed here.

To ensure that subjects did not hear more than one version of any sentence, two sets of experimental sentences were compiled; each target word occurred once in each set, and transition probability was counter-balanced across sets. For each set, there were three experimental tapes, making six conditions in all. The tapes were constructed as for Experiment 1, except that the blocks of 16 items each contained eight experimental, four warm up and four filler sentences. Timing marks were placed, and the offset of the marks from the target phonemes measured, as in Experiment 1.

The procedure was as for Experiment 1 except that there was only one version of the recognition test, corresponding to the first type of test used in Experiment 1.

3.1.2 Subjects

Forty-six subjects were tested, most of whom were members of the Applied Psychology Unit subject panel. All were native speakers of British English, and were between 18 and 41 years of age. They were paid for participating in the experiment, which lasted about 25 min. The results for one subject were lost due to equipment failure. Of the remaining 45 subjects at least six participated in each condition.

3.2. Results and discussion

The overall mean score on the recognition test was 60%. Sentences which had been heard as natural speech were correctly recognized no more often (64·5%) than sentences which had been heard in MITalk synthesized form (66·5%); the difference was not significant.

The number of missed targets was computed for natural vs. synthetic speech. On average subjects missed 1·13 targets in natural speech input and 1·91 targets in synthetic speech input; across subjects, this difference was significant ($t[44] = 3{\cdot}25$, $P = 0{\cdot}002$).

Mean RTs, adjusted for timing mark offset, again underwent separate analyses of variance with subjects and with items as random factors. The mean RT to natural speech was 505 msec, to synthetic speech 567 msec. This difference was significant in both analyses: $F1$ [1,44] = 21·63, $P < 0{\cdot}001$, $F2$ [1,20] = 25·32, $P < 0{\cdot}001$.

The results for the three further independent variables were:

(1) *Transition probability.* Targets on high transition probability words were responded to somewhat faster than targets on low transition probability words, but the difference did not reach significance and there was no interaction between transition probability and speech mode.

(2) *Position in sentence.* Targets on words which occurred late in the sentence were responded to significantly more rapidly than targets on words which occurred early in the sentence ($F1$ [1,44] = 41·77, $P < 0{\cdot}001$; $F2$ [1,20] = 51·01, $P < 0{\cdot}001$). However, the

advantage for late targets was 45 msec in the case of natural speech but more than twice as large—96 msec—in the case of synthetic speech. This interaction between sentence position and speech mode was also significant (*F1* [1,44] = 14·5, $P < 0{\cdot}001$; *F2* [1,20] = 5·12, $P < 0{\cdot}04$). *T*-tests across subjects revealed, however, that this interaction was due only to the difference in size of the effect in the two speech modes; late targets were detected significantly faster than early targets both in natural speech (t [44] = 3·39, $P < 0{\cdot}001$) and in synthetic speech (t [44] = 7·81, $P < 0{\cdot}001$).

(3) *Syllable stress.* RTs to targets on stressed vs. unstressed initial syllables were not significantly different either overall or in interaction with speech mode.

In contrast to Experiment 1, Experiment 2 showed a significant disadvantage for phoneme detection in synthetic as opposed to natural speech. The three further independent variables which we manipulated in the present study produced results similar to Experiment 1: there was in both experiments a weak but insignificant effect of transition probability, no effect of syllable stress, and a strong effect of sentence position (highly significant for both speech modes in each case, though the size of the effect for natural speech was reduced in Experiment 2). This consistency across experiments indicates that the two subject groups were processing the speech input in very much the same way. There is no obvious reason why the RT disadvantage for synthetic over natural speech should be insignificant in Experiment 1, but more than twice as large, and significant, in Experiment 2.

One problem with comparison of the two experiments is that both contained a third type of speech. The result of this was that certain aspects of experimental design were constrained differently in the two studies. In Experiment 1, the speech materials had originally been spoken spontaneously, whereas in Experiment 2 the materials were contrived sentences of a kind which is usual in sentence recognition experiments. In Experiment 1, subjects thus heard two-thirds natural speech and one-third synthetic speech input, while in Experiment 2 the synthetic speech formed two-thirds of the presented materials. However, on the face of it this latter difference might have been expected to produce the opposite effect, as Experiment 2 subjects received greater experience with synthetic input than subjects in Experiment 1; training is highly effective in improving listeners' processing of synthetic speech (Jenkins & Franklin, 1982; Schwab, Nusbaum & Pisoni, 1985; Greenspan, Nusbaum & Pisoni, 1988).

Another, minor, difference between the two experiments which might if effective have been expected to act in the opposite direction is the dialect of the speaker; in Experiment 1 the natural speech was produced by a British speaker, in Experiment 2 by an American. Since the synthetic speech had in both cases an American accent, and the listeners were in both cases British, the difference between natural and synthetic speech might if anything have been enhanced by this factor in Experiment 1. However, as British listeners hear American speech in the broadcast media daily, we consider that this factor is unlikely to have played a role.

A final minor difference between the two experiments was that the synthesizer employed in Experiment 1 was DECtalk and in Experiment 2 MITalk. There are certain differences between the two systems, notably the absence in DECtalk of the morphological decomposition module which enables MITalk to pronounce, for example, *scarcity* correctly (DECtalk's production being, roughly, *scar city*). However, these differences are irrelevant to the present studies since there were no such incorrect pronunciations in the output; as the basic principles of the two systems are identical, we again do not consider that this inter-experiment difference is likely to be the principal determinant of

the difference in results. Nevertheless, it is noticeable that Logan *et al.* (1989) found that the intelligibility of MITalk output, though at 93% very high, was significantly lower than that of DECtalk.

On balance, it is the nature of the speech materials which seem to be the primary candidate for explaining the difference between our two sets of results. In Experiment 3, listeners were presented with equal proportions of natural and synthetic speech input. This removed the possibility of any effect due to unequal proportions of input in the two modes. We manipulated the nature of the speech material by re-using material from the two previous experiments, again in equal proportion. Finally, we also compared both DECtalk and MITalk output.

4. Experiment 3

4.1. Method

4.1.1. Materials and procedure

The materials were chosen from those used in Experiments 1 and 2. From each experiment a set of sentences was selected to mimic the overall results from that experiment. Of the three factors investigated in both experiments (syllable stress, transition probability of the target word, and position in the sentence), the first two were omitted from the present experiment, since both previous experiments had found the same result for each. For sentence position, however, the results of Experiments 1 and 2 differed. Experiment 1 found a strong effect which was the same for both speech types, while Experiment 2 found a strong effect which was however significantly stronger for synthetic than for natural speech. In that both previous experiments found strong effects of this variable, including it here enables us also to check the consistency of the present results with the previous results.

This choice strongly constrained which sentences we chose from the two materials sets. Because the sentences of Experiment 1 were uncontrolled in their construction, i.e. spontaneously produced, variables such as sentence position were represented very unevenly. There were exactly six sentences with early targets. The mean RTs for these six sentences were closely similar for synthetic and natural speech; that is, the results for these sentences accurately mimicked the results for the experiment as a whole. We then selected six late-target sentences which also produced mean RTs which were closely similar for synthetic and natural speech, but which showed the strong position effect (i.e. the mean for the late-target sentences was about 100 msec shorter than the mean for the early-target ones). All twelve target words had stressed initial syllables and low transition probability. From Experiment 2, therefore, we selected six early-target and six late-target sentences (in fact, three pairs of each). All target words had stressed initial syllables and eight of the twelve had low transition probability (one high probability pair was included in the early and one in the late target set). The means for the chosen items accurately mimicked the means for Experiment 2 as a whole.

The 24 chosen sentences contained 3–5 instances of each of the 6 stop consonant targets /b,p,k,g,d,t/. A further 12 filler sentences were chosen from each of the previous experiments, allowing us to even up the occurrences of target phonemes to six of each. The sentences were again divided into two sets, A and B, with sentence source (Experiment 1 vs. 2) and position in sentence (early vs. late) counterbalanced across sets.

All sentences were recorded by a male native speaker of standard American English,

again selected for the similarity of his voice quality to that of the synthesizer. The recording conditions were the same as in the previous experiments. All sentences were also synthesized twice, once using DECtalk and once using MITalk. Timing marks were added to each version of each sentence in the same manner as for Experiments 1 and 2, and the offsets of the timing marks from the target phoneme were again measured.

The procedure was as for Experiments 1 and 2. The recognition test consisted of 24 items, half of which had occurred in the experiment and half of which had not. The latter were constructed of fragments from sentences which had occurred. Both the occurring sentences and the fragment-compiled sentences were balanced for membership of set A vs. B (thus, effectively, for whether a given subject had heard them in natural or synthetic form) and for source (Experiment 1 vs. 2).

4.1.2. Subjects

Forty-one subjects were tested, all native speakers of British English, and most of them from the Applied Psychology Unit subject panel. The age range was 18 to 40 years. The results for six subjects were lost due to equipment failure, and the results for a further three subjects who failed to score above 50% on the recognition test were discarded. Of the 32 remaining subjects, 16 heard DECtalk and 16 heard MITalk synthesized speech. Within each group of 16, eight heard set A sentences as natural speech and set B as synthesized, and eight vice versa; within each of these eight, four heard synthetic speech before natural speech and four vice versa.

4.2. Results and discussion

The average score on the recognition test (for those subjects whose data was analysed) was 77%. There was no significant difference between sentences which had been heard as natural or as synthesized speech, nor between sentences from Experiment 1 vs. 2.

Listeners again missed fewer targets in natural speech (1·16) than in synthetic speech (1·53), but in this case the difference was not statistically significant. However, the speech mode effect here interacted with presentation order ($F1$ [1,28] = 4·9, $P < 0·04$): subjects made more errors in the first half of the experiment than in the second, so that subjects who heard natural speech before synthetic speech made slightly more errors to natural than to synthetic speech, while subjects who heard synthetic speech before natural speech made many more errors to synthetic than to natural speech.

Natural speech (mean RT 468 msec) was responded to significantly faster than synthetic speech (mean RT 525 msec): $F1$ [1,28] = 25·32, $P < 0·001$; $F2$ [1,20] = 6·28, $P < 0·025$.

There was no main effect for order of presentation, but (in contrast to the two preceding experiments) the order variable did interact with speech mode: the RT advantage for natural over synthetic speech was 30 msec for subjects who heard synthetic speech first, but more than twice as large—83 msec—for subjects who heard natural speech first. This interaction was significant in both analyses: $F1$ [1,28] = 9·92, $P < 0·01$; $F2$ [1,20] = 5·45, $P < 0·04$. T-tests across subjects showed that the speech mode effect was highly significant for subjects who heard natural speech first (t [15] = 4·86, $P < 0·001$) but only marginal for subjects who heard synthetic speech first (t [15] = 1·97, $P < 0·07$).

As noted above, the order of presentation variable also interacted with speech mode in the analysis of missed targets. Putting these two effects together, we observe a speed-

accuracy tradeoff: subjects tended to respond more rapidly, but miss more targets, in whichever speech mode they heard first; in the second half of the experiment they missed fewer targets, but on average responded more slowly.

Results for the three further variables were:

(1) *Synthesizer.* There was no significant difference in RT to DECtalk (532 msec) vs. MITalk (518 msec), and nor was the size of the RT advantage for natural over synthetic speech significantly different for DECtalk vs. MITalk.

(2) *Sentence source.* There was no significant difference in RT to sentences from Experiments 1 and 2 respectively. However, there was a clear replication of the results of the previous two experiments in the way the two types of sentences were responded to in natural vs. synthetic speech. The RT advantage for natural over synthetic speech was 30 msec for Experiment 1 sentences but 84 msec for Experiment 2 sentences. The interaction between sentence source and speech mode was significant in the analysis by subjects ($F1$ [1,28] = 5·07, $P < 0·04$) but failed to reach significance in the analysis by items. T-tests across subjects revealed that the speech mode effect was highly significant for Experiment 2 sentences (t [31] = 4·41, $P < 0·001$) but marginal for Experiment 1 sentences (t [31] = 1·73, $P < 0·095$).

(3) *Position in the sentence.* Late targets were detected faster than early targets: $F1$ [1,28] = 117·91, $P < 0·001$; $F2$ [1,20] = 23·76, $P < 0·001$. This effect did not, however, interact with speech mode or with any other variable. The RT advantage for late targets was 125 msec for natural speech and 114 msec for synthetic speech; t-tests across subjects revealed that both differences were significant (t [31] = 8·54, $P < 0·001$ for natural speech, t [31] = 5·98, $P < 0·001$ for synthetic speech). The three-way interaction between speech mode, sentence source and position in the sentence did not reach significance, but again the pattern of responses reflected that of the two preceding studies: for the sentences from Experiment 1 there was a larger position effect with natural than with synthetic speech, while for the sentences from Experiment 2 there was a larger position effect with synthetic than with natural speech.

This third experiment has shown that the difference in findings of Experiments 1 and 2 was not due to type of synthesizer, or to choice of sentences. In the present experiment subjects' response patterns were very similar to those in the preceding experiments. However, close analysis of the response patterns shows evidence of a speed–accuracy tradeoff, i.e. a shift in subjects' response criterion.

5. General discussion

Taken together, the three experiments warrant several general conclusions about the processing of synthetic speech at the phonetic level, and the way this processing can be investigated.

First, it is clear that the synthetic speech which we presented to our listeners was highly intelligible. In all three experiments, listeners performed as well on the recognition test with sentences which they had heard in synthesized form as with sentences which they had heard in natural form.

Moreover, all the independent variables that previous phoneme detection studies prompted us to observe here revealed response patterns which were the same for natural and for synthetic speech. This further confirms the conclusions of previous researchers (e.g. Pisoni, 1981; Pisoni, Manous & Dedina, 1987) that there are no real differences in the way listeners process synthetic in comparison to natural speech.

Nevertheless, the high intelligibility did not prevent a consistent disadvantage for synthetic speech as measured by RT to detect the phoneme targets. In all three experiments (though to a statistically significant degree only in two) subjects responded more rapidly to targets in natural speech than in synthetic speech. Thus the phoneme detection task has proved to be a very useful tool for investigating the processing of synthetic speech. We have replicated the finding of Pisoni, Manous & Dedina (1987) that RT measures are capable of demonstrating processing disadvantages for synthetic speech even in the absence of intelligibility differences. Phoneme detection therefore joins other RT measures in the repertoire available to investigators who will need ever more subtle tools to assess the relative processing ease of ever higher quality synthesizer output. Moreover, it provides a tool which is dedicated specifically to measurement of phoneme perception.

However, it is clear that characteristics of the experimental materials need to be carefully considered in the design of phoneme detection experiments. In Experiment 1, using sentences which had originally been produced as spontaneous speech, the RT difference between natural and synthetic speech did not reach significance. Had the same result appeared in Experiment 2, we might have felt justified in making extravagant claims about the ease with which phonemic representations can be constructed from synthetic speech input. However, Experiment 2 produced a highly significant RT disadvantage for synthetic speech. Only the results of Experiment 3 allowed this contradiction to be resolved.

One key aspect of the Experiment 3 results holds the answer: the fact that subjects clearly showed a speed–accuracy tradeoff across the experiment. In whichever type of speech material they heard first, our subjects responded quite rapidly but made relatively more response errors; in whichever type of material they heard second, their responses were more accurate but slower.

Apparently, therefore, subjects shifted their response criterion across the experiment. What can have led them to do this? Obviously it cannot be the presence of synthetic as well as naturally spoken input, because the criterion shift happened whichever type of speech material they heard first. We suggest that it was the nature of the sentence materials which led, albeit indirectly, to the response criterion adjustment. Recall that in Experiment 3 half of the materials were taken from Experiment 1 and half from Experiment 2. Thus it would only gradually become apparent to the subjects that some sentences contained ambiguous pronouns, unclear diectic expressions, and the like, and hence would require inferential work for their interpretation. As this characteristic of the materials became clear, it caused subjects to shift towards slower, more careful responding; this shift then showed up as slower RTs but fewer missed targets in the second half of the experiment. Slower and more careful responding is likely to wipe out the response time disadvantage of synthetic speech if this disadvantage is due, as argued by Pisoni, Nusbaum and Greene (1985) among others, to difficulties in the earliest stages of perceptual analysis.

In Experiment 1, therefore, the lack of significance in the response time disadvantage for synthetic speech can be explained as slow and careful responding throughout the experiment. Since *all* the materials of Experiment 1 were of the spontaneous, i.e. contextually dependent, type, subjects would adopt a careful response criterion at an early stage of the experiment. Consistent with this account is the fact that the overall RTs in Experiment 1 were the slowest of all the three experiments (though it is of course impossible to make much of this given that quite separate subject groups were tested in

the three experiments). Also consistent is the lack of order of presentation effects in Experiment 1, and the fact that there were not significantly more missed targets in either natural or synthetic speech. Indeed, the number of missed targets was, just as this account would predict, lower in Experiment 1 than in either of the later experiments. Thus although Experiment 1 subjects may well have experienced difficulty in constructing phonetic representations from synthetic speech, the response criterion which they adopted resulted in responses being issued at a stage when such initial processing difficulties had been overcome.

In Experiment 2, in contrast, subjects were presented with contextually independent sentences, and adopted a response criterion which laid more emphasis on speed than the criterion encouraged in Experiment 1. The result of this was that targets in synthetic speech were responded to significantly more slowly than targets in natural speech. Again the materials were constant in nature, and there was no criterion shift across the experiment; consistent with this is the lack of order effects. Note, however, that the analysis of missing targets in Experiment 2 produced a significant difference: more targets were missed from synthetic than from natural speech materials. The mean number of missed targets was also higher in Experiment 2 than in either of the other two experiments. This is exactly as would be predicted by an account in which subjects have adopted a response criterion which produced faster but less careful responding.

If the aim of a phoneme detection experiment is to investigate processing at the phonetic encoding stage, therefore, the nature of the speech materials is of crucial importance in the experimental design. Note that this is not a simple effect. It is not, for instance, the case that the materials of Experiment 1 were intrinsically harder than those of Experiment 2. Analysis of the materials showed that the mean frequency of occurrence of the content words in Experiment 1 was higher than in Experiment 2, and that Experiment 1 contained a larger proportion of function words – thus on the face of it the materials of Experiment 1 would appear easier. In fact, the lack of a main effect of sentence source in Experiment 3 indicates that the two materials sets were equally easy to process. However, subjects were encouraged by the Experiment 1 materials to adopt a more careful response criterion; we assume that the reason for this was the contextually dependent nature of these materials.

There are a number of ways in which subjects' responding in phoneme detection experiments can be altered by the nature of the stimulus materials; Cutler *et al.* (1987) discuss these issues in greater depth. For phoneme detection to be useful as a tool for investigation of phonetic processing during the comprehension of synthetic speech, it is important that subjects be encouraged to adopt a response criterion which emphasizes rapid responding. When our subjects did so, a response time disadvantage for synthetic speech appears. Despite a very high level of intelligibility for this particular high-quality synthetic speech (Logan *et al.*, 1989), the present phoneme detection results indicate that it does not yet rival natural speech in the ease with which listeners can extract phonetic information. The phoneme detection task offers a new tool for investigating this particular aspect of the perception of synthetic speech.

Acknowledgements

The authors' names are listed in reverse alphabetical order. The experiments reported in this paper formed part of thesis projects for the degree of M.Phil. in Computer Speech and Language Processing at Cambridge University, and were conducted by the second, third, and first author

respectively under the supervision of the fourth author. Financial support is acknowledged from the Science and Engineering Research Council (AJN), British Gas (GM) and GPT Telecommunications (JD). The authors are grateful to Sally Butterfield, Mark Coulson, John Culling, Ian Nimmo-Smith and Dennis Norris for assistance. AJN is now at the Department of Psychology, Birkbeck College, University of London; GM is with the Centre for Petroleum and Mineral Law Studies, University of Dundee; and JD is with HB Technologies Ltd. (Newton Aycliffe, UK).

References

Allen, J., Hunnicutt, S. & Klatt, D. H. (1987). *From Text to Speech: MITalk System*, Cambridge University Press, Cambridge.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55–60.

Cutler, A. & Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition*, **7**, 49–59.

Cutler, A. & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, **20**, 1–10.

Cutler, A., Mehler, J., Norris, D. & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, **19**, 141–177.

Cutler, A. & Norris, D. (1979). Monitoring sentence comprehension. In *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett* (Cooper, W. E. and Walker, E. C. T., eds), pp. 113—134. Erlbaum, Hillsdale, New Jersey.

Dell, G. S. & Newman, J. E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behaviour*, **19**, 609–623.

Digital Equipment Corporation. (1983). *DECtalk DTC01 Owner's Manual*. Merrimack, New Hampshire.

Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behaviour*, **8**, 457–462.

Foss, D. J. & Gernsbacher, M. A. (1983). Cracking the dual code: Toward a unitary model of phonetic identification. *Journal of Verbal Learning and Verbal Behaviour*, **22**, 609–632.

Greenspan, S. L., Nusbaum, H. C. & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **14**, 421–433.

Hoover, J., Reichle, J., Van Tasell, D. & Cole, D. (1987). The intelligibility of synthesised speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*, **30**, 425–431.

Jenkins, J. J. & Franklin, L. D. (1982). Recall of passages of synthetic speech. *Bulletin of the Psychonomic Society*, **20**, 203–206.

Kučera, H. & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*, Brown University Press, Providence, RI.

Larkey, L. S. & Danly, M. (1983). Fundamental frequency and sentence comprehension. *MIT Speech Communication Group Working Papers*, **2**, 25–39.

Lee, L. & Nusbaum, H. C. (1989). The effects of perceptual learning on capacity demands for recognising synthetic speech. Paper presented to the 117th meeting, Acoustical Society of America, Syracuse.

Levelt, W. J. M. (1978). A survey of studies in sentence perception. In *Studies in the Perception of Language*, (Levelt, W. J. M. & Flores d'Arcais, G. B., eds), pp. 1–74. Wiley, New York.

Logan, J. S., Greene, B. G. & Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566–581.

Luce, P. A., Feustel, T. C. & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25**, 17–32.

Mehler, J. & Segui, J. (1987). English and French speech processing. *The Psychophysics of Speech Perception*, (Schouten, M. E. H., ed.), pp. 405–418. Martinus Nijhoff, Dordrecht.

Mehler, J., Segui, J. & Carey, P. W. (1978). Tails of words: Monitoring ambiguity. *Journal of Verbal Learning and Verbal Behaviour*, **17**, 29–35.

Mehta, G. & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, **31**, 135–156.

Morton, J. (1967). Population norms for sentence completion. Unpublished manuscript, Applied Psychology Unit, Cambridge.

Morton, J. & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behaviour*, **15**, 43–51.

Newman, J. E. & Dell, G. S. (1978). The phonological nature of phoneme monitoring: A critique of some ambiguity studies. *Journal of Verbal Learning and Verbal Behaviour*, **17**, 359–374.

Norris, D. G. (1984). A computer-based programmable tachistoscope for non-programmers. *Behavior Research Methods, Instrumentation and Computers*, **16**, 25–27.

Nusbaum, H. C., Dedina, M. J. & Pisoni, D. B. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. In *Research on Speech Perception: Progress Report 10*, Speech Research Laboratory, Indiana University; 409–422.
Pisoni, D. B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. Paper presented to the 102nd meeting, Acoustical Society of America, Miami.
Pisoni, D. B., Manous, L. M. & Dedina, M. J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech & Language*, **2**, 303–320.
Pisoni, D. B., Nusbaum, H. C. & Greene, B. G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **11**, 1665–1676.
Ralston, J. V., Pisoni, D. B., Lively, S. E., Greene, B. G. & Mullenix, J. W. (1991). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*, **33**, 471–491.
Ralston, J. V., Pisoni, D. B. & Mullenix, J. W. (in press). Comprehension of synthetic speech produced by rule. In *Behavioral Aspects of Speech Technology: Theory and Applications*, (Bennett, R., Syrdal, A. M. & Greenspan, S., eds), Elsevier: New York.
Schwab, E. C., Nusbaum, H. C. & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, **27**, 395–408.
Shields, J. L., McHugh, A. & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, **102**, 250–255.
Silverman, K. E. A. (1987). *The Structure and Processing of Fundamental Frequency Contours*, PhD Thesis, University of Cambridge.
Taft, L. (1984). *Prosodic Constraints and Lexical Parsing Strategies*, Ph.D. Thesis, University of Massachusetts.
Terken, J. & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics*, **16**, 453–457.
Waterworth, J. A. & Holmes, W. J. (1986). Understanding machine speech. *Current Psychological Research and Reviews*, **5**, 228–245.

Appendix

Experimental Sentences

Each sentence is preceded by its phoneme target. Asterisked sentences were also used in Experiment 3. Experiment 2 sentences occurred in two versions, with target-bearing words of high and low (bracketed) transition probability.

Experiment 1

/d/ I've always heard of Cambridge described as such I think.
/b/ The most important thing is to buy the right make. (*)
/g/ Spock gradually learns to swear. (*)
/p/ They have always allowed us to publish thus far. (*)
/t/ I think you shouldn't buy either of them for the time being because at the moment they're incompatible with each other.
/k/ Apparently there was a considerable exodus around 1979.
/d/ It makes quite a considerable difference to be running behind someone else.
/b/ One of the things Dolby does is that it boosts up the high frequency.
/g/ Just a set of words that had the words in groups of six or seven.
/p/ The best people in the world are just under 45 seconds. (*)
/t/ It only ran for three years, when it was on television, and they haven't made any since. (*)
/k/ Have you got a very big record collection?
/d/ Everything seems to be very democratic here. (*)
/b/ You'd have to turn it from digital back into analog.
/g/ The research which is going on here is pretty fundamental.

/p/ The intrinsic idea of having only a single power supply and running everything off it is a good idea. (*)
/t/ I wouldn't be surprised if there was quite a big effect of tactical voting. (*)
/k/ All people who call themselves psychiatrists are in fact medics. (*)
/d/ When you're doing it for yourself, there isn't really much of a batch. (*)
/b/ They have the right to stop us publishing I believe.
/g/ They were very psychological, I grant you that.
/p/ Individual researchers have their individual projects and get on with them.
/t/ That makes the plates move together and apart again, and pushes the air back and forth.
/k/ You'll have to accept that something better may come along which you won't be able to use.
/d/ She stands a better chance of defeating the Conservative.
/b/ If you've seen them in the shops, you'll see that they're very big and very flat. (*)
/g/ I don't think it's a very good trend towards the American way of doing things.
/p/ The electric charges on the two plates are varied by the amplifier.
/t/ What tests can be done on attention? (*)
/k/ Daley Thompson is just over 45 seconds, which is amazing considering he has nine other events to do.

Experiment 2

Stressed, early:
/k/ The mother would kiss (*) (kick) the children often, the social worker said.
/k/ The striker kicked (*) (kissed) the ball in triumph.
/g/ At the art gallery (gathering (*)) there was a lot of talk about faking.
/g/ In the large teeming gathering (gallery (*)) it was impossible to find anyone.
/p/ The wall fell on a passerby (pacifist (*)), who was injured quite severely.
/p/ "Ban the bomb!" cried the pacifist (passerby (*)) to the shoppers.

Stressed, late:
/g/ While he was preparing the soup the chef crushed some garlic (*) (garbage) too.
/g/ The cookery teacher told the class to throw out the garbage (*) (garlic) right away.
/b/ The chairman told the speaker to keep it brief (bright (*)) as the audience was tired.
/b/ Despite her sadness her speech of welcome was bright (brief (*)) and cheerful.
/d/ The lawyers opposing the move said it would be detrimental (devastating (*)) to their clients' prospects.
/d/ She told the psychiatrist that his unfaithfulness had had a devastating (detrimental (*)) effect on her work.

Unstressed, Early:
/d/ The wearying delay (defeat) had left them all extremely tired.
/d/ The team's latest defeat (delay) had ruined their confidence.
/b/ Monkeys ate the bananas (belongings) which the campers had left near the tent.
/b/ Carrying their belongings (bananas) the plantation workers made their way down the road.

/p/ The official permission (petition) he sought never materialized.
/p/ The complainants' petition (permission) to drop the case angered the judge.

Unstressed, Late:
/d/ His previous experience made him distrust (disturb) the neighbours.
/d/ He hoped the new neighbours would not disturb (distrust) them.
/p/ The managers discussed which junior to promote (prefer) for the job.
/p/ The salesmen voted on which scheme they would prefer (promote) next.
/k/ He had to take a later train after he missed his connection (collection) at Peterborough.
/k/ At the end of the service the minister made a collection (connection) which surprised everyone.