Perception & Psychophysics
1978, 24 (6), 496-500

Syllabic and phonemic representations for short-term memory of speech stimuli

PETER HOWELL
University College London, Gower Street, London WC1E 6BT, England

An experiment is reported which uses a same-different matching paradigm in which subjects are required to indicate whether the consonants of a pair of consonant-diphthong syllables are the same or different. The question addressed is the operation of two hypothesized processes in the perception of speech sounds. The auditory level is shown to hold stimulus information for a brief period of time and be sensitive to allophonic variations within a stimulus. Moreover, matching at this level takes place by identity of the syllables rather than of the separate phoneme segments. The phonemic level is impaired when the diphthong segments of the pair leads to a contradictory match to that of the consonants of the pair, even though only the consonants are relevant to the matching decision.

Intelligible speech patterns can be synthetically produced by representing the variations over time of the low-frequency resonances (formants) of the vocal tract. The movement of these resonances reflects the change in the positioning of the articulators. Sounds produced in this manner have been extensively used to determine the cues responsible for speech perception. For example, speech sounds can be synthesized which simulate the movement of the tongue to cue the different places of articulation for the voiced stop consonants /b/, /d/, and /g/. Considering two-formant patterns, changing the place of articulation of a consonant-vowel syllable is reflected in a change in the starting position of the second formant transition. It starts at a low value for /b/, is intermediate for /d/ and is high for /g/ when preceding the vowel /a/. The vowel portion of the spectrogram is constant for the rest of the syllable and has the same formant frequencies for /b/, /d/, and /g/. The formant frequencies of the consonant gradually move to the starting position of the vowel formants.

Sounds intermediate between the formant frequencies appropriate for a particular phoneme can also be synthesized. When these are presented to subjects to classify into one of the voiced stop categories, /b/ is heard for a range of second-formant starting frequencies, then there is an abrupt change to hearing /d/, and finally, a change to /g/. The points of abrupt change are called phoneme boundaries, and the variations within a phoneme which are not perceived by a listener are termed intrinsic allophonic variations. The sudden change in phoneme report and the fact that subjects can discriminate between stimuli better when they are on opposite sides of the phoneme boundary than when they are on the same side led investigators from the Haskins Laboratory to postulate the existence of a phonemic mechanism not sensitive to variations within a phoneme which is probably localized in the left hemisphere (Liberman & Studdert-Kennedy, Note 1).

Subsequent experimental and theoretical considerations emphasized the need to postulate a memory which holds a literal representation of the stimulus for a brief period of time before or while it is being analyzed into its phonemic class (Darwin & Baddeley, 1974; Fujisaki & Kawashima, 1968). For example, Darwin and Baddeley (1974) argued that an auditory memory susceptible to degradation over time would have the fine detail responsible for the distinction between, say, the stop consonants /b, d, g/ destroyed before that between the more discriminable vowels /i, I, u/. Thus there may be more opportunity to use the information in auditory memory for decisions about these vowel stimuli than for the stop consonants, and it might be expected, as has been reported, that there will be differences in the type of categorization function for these classes of speech sounds (Liberman, Harris, Hoffman, & Griffith, 1957).

In order to obtain evidence for auditory memory for stop consonants, the application of more sensitive paradigms was called for. One such paradigm was first employed by Pisoni and Tash (1974). In this paradigm, subjects are required to indicate if the second of two sequentially presented syllables is the same or different. The time for subjects to make this decision is measured. For the same pairs, the two stimuli can be identically the same or allophonic variations of the same phonemes (nonidentical same). Subjects should respond faster when the pairs are identical if this information is available in auditory memory and decisions can be made on this informa-

The author is in receipt of an SRC studentship and is supervised by Professor R. J. Audley. Part of this work was reported at the 20th Tagung Experimentell Arbeitender Psychologen at Marburg, March 1978.

tion (this is referred to as the same matching advantage). This result has usually been found (Howell & Darwin, 1977; Pisoni & Tash, 1974; Eimas & Miller, Note 2). In addition, varying the separation between the first and second stimulus causes the same matching advantage to disappear (Howell & Darwin, 1977; Eimas & Miller, Note 2). This finding demonstrates that the memory is of short duration. In their initial report, Pisoni and Tash (1974) reasoned that different decisions might be made at an auditory level too, but other authors have criticized this line of thinking (Howell & Darwin, 1977; Eimas & Miller, Note 2; Repp, Note 3). In the present experiment, conclusions about auditory memory are only based on the same matching advantage.

The evidence reviewed to support a mechanism that identifies the phonemes does not require that both the auditory and phonemic levels hold representations of stimuli corresponding to phonemes. Indeed, one line of argument suggests that a syllabic representation must be available at *some* point. Thus, the information about the separate phonemes is interleaved among syllable-sized units. This phenomenon, known as coarticulation, can be appreciated from the discussion of the formant frequencies of the stop consonants /b, d, g/, where it may be noted that the direction of the transitions is determined by the formants of the vowel. The information about the phonemes is spread throughout the syllable, and it would seem necessary for this information to be available at an early stage in processing. Syllables might, then, be held in auditory memory.

The question addressed in the present study is whether the auditory level is able to distinguish the separate phonemes or whether decisions at this level are made between the overall identity of the pair of syllables. One way to do this is to change the phoneme segment which subjects are not required to classify (the irrelevant segment) and see whether the same matching advantage still exists for the classified (relevant) segment. If it does occur, then auditory memory holds information about the separate phonemes. A problem in designing the experiment is that if the formant frequencies are changed to produce a different vowel, the frequency components of the consonant changes, too, as a result of coarticulation effects. Thus, a way of varying the irrelevant segment is needed which need not necessarily lead to changes in the relevant segment. This may be achieved by having the formant frequencies converge onto the same frequencies for the irrelevant segment. By employing diphthongs as the irrelevant segment, it can be arranged for these to diverge from the same starting frequencies and hence allow variations in each phoneme segment independently. Diphthongs are produced when two vowels are adjacent to each other, and inspection of their spectrograms shows that the formant frequencies move gradually from the position appropriate for one vowel to those appropriate for the other. The diphthongs /au/ and /eI/ were used as the irrelevant segment where the frequency of the second formant of /au/ drops over time while that for /eI/ rises.

If the auditory level segments the stimulus into its separate phonemes, the same matching advantage is expected whether or not the diphthong of the second stimulus is identically the same as, nonidentically the same as, or different from that of the first stimulus. If, on the other hand, the auditory level operates on the overall identity of the syllables, then a nonidentical same or different diphthong should remove the same matching advantage.

In addition to this question, the separation between the first and second stimulus was varied as in the experiments of Howell and Darwin (1977) and Eimas and Miller (Note 2) in order to further substantiate the finding that the information in auditory memory is subject to rapid decay.

METHOD

Subjects

Eight subjects aged between 18 and 23 and all right-handed attended for three sessions of about 2 h each. They were paid at the rate of 60 p/h.

Stimuli

The consonant portion of each stimulus lasted 40 msec and the diphthong, 100 msec. For each stimulus, the first formant rose linearly from 175 to 700 Hz over 40 msec. For the diphthong portion, it then decreased linearly to 380 Hz for both diphthongs. The second-formant starting frequencies were 1,206 and 1,283 Hz (allophonic variations of /b/), rising to 1,500 Hz, or 1,752 and 1,829 Hz (allophonic variations of /d/), decreasing to 1,500 Hz again.

The second-formant transitions of the diphthongs decreased to 1,020 or 980 Hz (allophonic variations of /au/) or increased to 1,980 or 1,940 Hz (allophonic variations of /eI/).

Each of the consonants could be paired with each of the diphthongs, giving 16 stimuli in all. These were computed on a software parallel-formant synthesizer running on a PDP-12 computer. The stimuli were output at 8 kHz and filtered at 3.5 kHz before recording on a Revox tape recorder.

The same-different matching tape was recorded by pairing each of the 16 stimuli as initial stimulus with each of the stimuli as the final stimulus once, giving 256 matching trials in all. These were then randomized, and three of the randomizations were recorded with a 250-, 500-, or 1,000-msec pause between them. Between each matching trial, there was a pause of 3 sec.

Procedure

Each of the 16 stimuli was presented four times at the beginning of each session from a prerecorded randomization recorded on tape. The subject had to classify the stimuli as bow, dow, bay, or day. The subject then received the same-different matching condition. He was required to indicate, by means of a pair of response keys, whether the syllables had the same or a different consonant. He did this as quickly and as accurately as possible, and his response and the time for this decision was recorded on a PDP-12 computer.

The subject received each of the three randomizations within each session but in different orders across sessions. In addition, order of the randomizations was counterbalanced across subjects.

RESULTS

The results of the identification test showed that no stimulus was assigned to a consonant class other than expected on more than 2.1% of the time (diphthong classification gave the expected class on more than 99% of the time). For the matching conditions, the first session was dropped from data analysis. The mean correct reaction time and number of errors for the experimental conditions are presented in Figures 1 and 2, respectively.

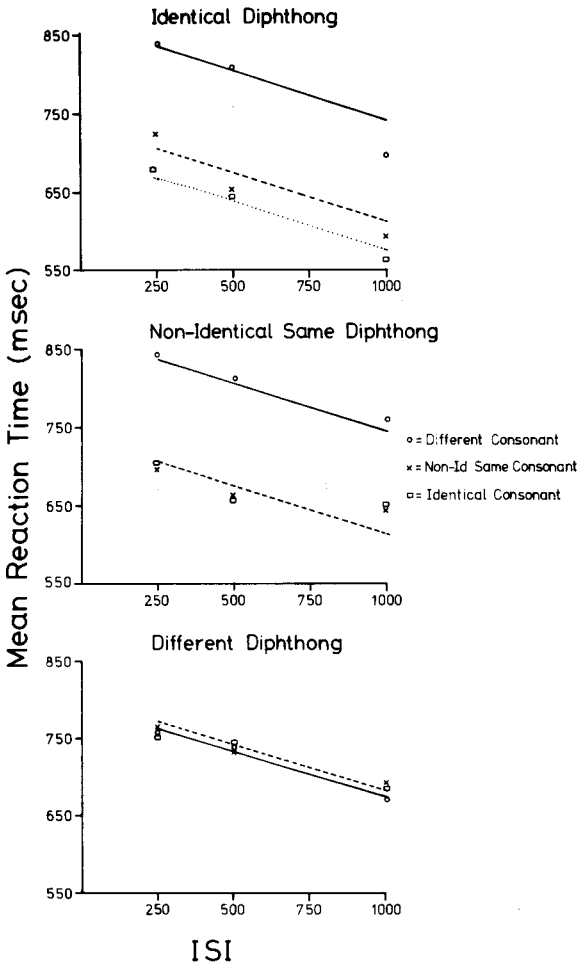

Figure 1. Mean correct reaction time for each of the diphthong conditions. The symbols for the consonant conditions are given in the inset. The abscissa is the separation between the two syllables. The dotted line represents the fit of Equation 1, the dashed line of Equation 2, and the solid line of Equation 3. These represent the models for the auditory, phonemic when the stimuli are the same, and phonemic when the stimuli are different processes, respectively. Except in the identical diphthong condition, auditory information cannot be used and the appropriate curve is that for the phonemically same process.

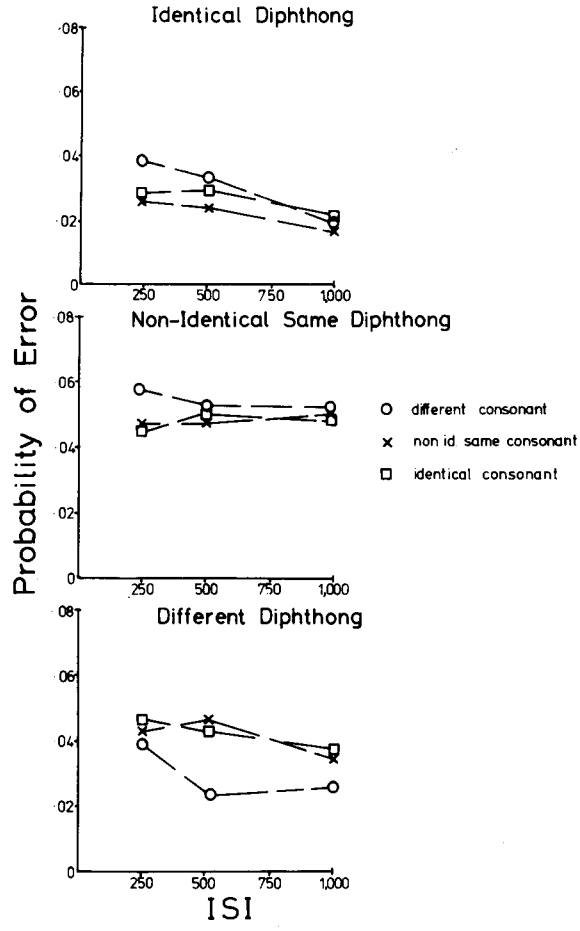

Figure 2. Error rates for each of the diphthong conditions. The parameters of the connected points are the consonant conditions, and the abscissa is the separation between the two syllables.

Each of the diphthong conditions was analyzed by an analysis of variance which distinguished subjects (eight levels), consonant condition (identical, nonidentical same, or different), and interstimulus interval (ISI: 250, 500, or 1,000 msec). For the identical diphthong condition, there was a main effect of consonant matching condition [$F(2,14) = 11.6$, $p < .01$] and of ISI [$F(2,14) = 5.6$, $p < .05$], as well as an interaction between them [$F(4,28) = 3.3$, $p < .05$]. Planned comparisons of the consonant condition effect showed that the difference between the identically same consonant and the nonidentically same consonant [$F(1,14) = 5.1$, $p < .05$] and the difference between the different consonant and the nonidentically same consonant [$F(1,14) = 6.3$, $p < .05$] were significant. Planned comparisons of the Consonant by ISI interaction revealed that only the identically same to nonidentically same consonant difference interacted with ISI [$F(1,28) = 4.9$, $p < .05$].

For the nonidentically same diphthong, there was a main effect of consonant condition [$F(2,14) = 10.8$, $p < .01$], which was due to a difference between the

nonidentically same consonant and the different consonant [$F(1,14) = 6.9$, $p < .05$], with nonidentical same consonants being faster than the different consonants. There was also a main effect of ISI [$F(2,14) = 7.3$, $p < .01$].

For the different diphthong condition, there was only an effect of ISI [$F(2,14) = 6.1$, $p < .05$].

The corresponding analyses were performed on the error rates after arcsin root transformation. The only effect to reach significance was consonant condition for the different diphthongs [$F(2,14) = 5.6$, $p < .05$]. This was due to the different consonant (which, in this condition, also had a different diphthong), giving a lower error rate than the other consonant conditions.

DISCUSSION

The first conclusion that can be drawn from the data is that the same matching advantage only occurs when the consonant *and* diphthong of the second stimulus are identical in terms of the consonant *and* diphthong to those of the first stimulus. Since it is only in the condition where the syllables are identical that the same matching advantage is found, the auditory level cannot discriminate phonemic segments; decisions at this level are based on the overall identity of the syllables.

Further support for this conclusion is suggested from the results of an unpublished experiment by Darwin and Howell. They employed the same-different matching paradigm and timed subjects to indicate whether the stop consonant of the second stimulus was the same as or different from that of the first. The consonants of the pairs could be identically the same or nonidentically the same as in the present experiment. The vowel employed as the irrelevant segment (/ae/) could be the same length or different to that of the first (both short, both long, short then long, or long then short). The sequences were arranged so that the stimulus onset asynchrony was always the same. The same matching advantage was found only when the second stimulus was the same length as that of the first (i.e., both long or both short). The variation in the irrelevant segment (vowel length) disrupted the same matching advantage as in the present experiment and substantiates the conclusion that the auditory level operates on the identity of the syllables of the two stimuli.

Inspection of the nonidentical same consonant and different consonant response profiles reveals some interesting differences. The separation between these stimuli is large in the identical and nonidentical same diphthong conditions, with the nonidentical same consonant being faster than the different consonant. There is no difference between these consonants in the different diphthong condition, but here the absolute magnitude of the response times lies between those of the nonidentical same and different consonant conditions of the other diphthong conditions. These differences are surprising when it is remembered that all of these responses are reasoned to be based on the output of the phonemic mechanism. If equal processing time was required for each type of stimulus at this level, then the results which might have been expected are those which were obtained for the different diphthong condition assuming that the time for a same or a different response is approximately equal. The results can be explained on the assumption that there is inhibition of a same response when the irrelevant segment is different and, conversely, inhibition of a different response when the irrelevant segment is the same. To see how well these assumptions explain the data, the following equations were fit, using an iterative technique:

$$RT_{identical} = t_s - w(ISI) - t_{aud}e^{-\lambda(ISI)} \quad (1)$$

$$RT_{nonidentical} = t_s - w(ISI) + n_d tsp \quad (2)$$

$$RT_{different} = t_d - w(ISI) + n_s tdp, \quad (3)$$

where t_s = processing time required for same pairs, t_d = processing time required for different pairs, tsp = inhibition of same processing caused by a different phoneme, tdp = inhibition of different processing caused by a same phoneme, w = slope constant of ISI effect, t_{aud} = processing time of auditory level, λ = decay constant of information from auditory level, n_d = number of phonemes different for a "same" pair, and n_s = number of phonemes same for a "different" pair.

The lines in Figure 1 are the obtained solutions. The weighting factor, w, represents the decrease in response time with increasing ISI which arises presumably as a result of preparation. The exponential function represents the decay from auditory memory. The processing time for same and different pairs (t_s, t_d) was not assumed to be equal (Egeth, 1966), nor was the amount of inhibition caused by the type of phoneme present in the syllable not corresponding to the required response (tsp, tdp). Estimating seven parameters from the 27 data points left a residual RMSD of 12 msec.

Note that although no direct evidence has been offered in favor of phoneme processing at perceptual levels beyond the auditory store, if only syllables were represented at the "phonemic" level, it would still be necessary to invoke phonemic processes because, for different responses (where the stimuli are always different syllables), there are different response times, depending on the number of phoneme segments by which the stimuli differ.

The interaction between ISI and the same matching advantage for identical diphthongs indicates that information in this store lasts for only a short time, since the advantage disappears as the separation between the two stimuli is increased. Individual t tests showed that the same matching advantage for identical diphthongs was significant at 250 msec ISI and at 500 msec ISI but not at 1,000 msec. This substantiates the findings of Howell and Darwin (1977), where a same matching advantage was found at 50 and 200 msec ISI but not at 800 (see also, Eimas & Miller, Note 2).

Given that the auditory level operates on syllable-sized units, the question still exists as to how processing progresses from these units to sentences. This issue remains unresolved and is made more complicated in the light of the finding that the properties of auditory memory, as revealed by other paradigms, show a susceptibility to overwriting from information which arrives after the stimulus (Crowder, 1972; Crowder & Morton, 1969). The problem of how perceptual systems extract continuous information from the physical world is not confined to the auditory modality; for example, why is the visual image not blurred when the eye moves?

In summary, the present experiment has shown that the auditory level operates on syllable-sized units. The irrelevant phoneme of the second stimulus impairs the phonemic level when it gives rise to a conflicting outcome to that of the relevant phoneme. Further support for the short-term storage of the auditory level is presented.

REFERENCE NOTES

1. Liberman, A. M., & Studdert-Kennedy, M. *Phonetic perception*. Haskins Laboratories Status Reports on Speech Research, 1977, **SR-50**, 21-60.

2. Eimas, P. D., & Miller, J. *Auditory memory and the processing of speech*. Brown University Progress Report No. 3, Providence, Rhode Island: W. S. Hunter Laboratory of Psychology, 1975.

3. Repp, B. H. *Posner's paradigm and categorical perception: A negative study*. Haskin's Laboratories Status Report on Speech Research, 1976, **SR-45/46**, 153-161.

REFERENCES

CROWDER, R. G. Visual and auditory memory. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye*. Cambridge: MIT Press, 1972.

CROWDER, R. G., & MORTON, J. Precategorical acoustic storage (PAS). *Perception & Psychophysics*, 1969, 5, 365-373.

DARWIN, C. J., & BADDELEY, A. D. Acoustic memory and the perception of speech. *Cognitive Psychology*, 1974, 6, 41-60.

EGETH, H. E. Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics*, 1966, 1, 245-252.

FUJISAKI, H., & KAWASHIMA, T. The influence of various factors on the identification and discrimination of synthetic speech sounds. *Reports of the 6th International Congress on Acoustics*, Tokyo, Japan, 1968.

HOWELL, P., & DARWIN, C. J. Some properties of auditory memory for rapid formant transitions. *Memory & Cognition*, 1977, 5, 700-708.

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S., & GRIFFITH, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, 54, 358-368.

PISONI, D. B., & TASH, J. Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 1974, 15, 285-290.

(Received for publication January 17, 1978; revision accepted September 15, 1978.)