Journal of Phonetics (1990) **18**, 229–254

Contrast and normalization in vowel perception

Keith Johnson

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405 U.S.A.

Received 2nd November 1989, and in revised form 9th February 1990

The work reported in this paper is an attempt to understand vowel normalization better by investigating the relationship between vowel normalization and vowel contrast. In the first experiment, vowels from a "hood"–"hud" continuum were presented at two levels of fundamental frequency (F_0) using two types of presentation. In one condition, tokens were blocked by F_0. In the other, tokens with different F_0 levels were randomly intermixed with each other (as in the typical F_0 normalization experiment). In the mixed presentation, subjects identified the high F_0 items most often as "hood" and the low F_0 items most often as "hud". In the blocked condition, there was no reliable difference between the high and low F_0 continua. This pattern of results suggests that a contrast effect is at work. Therefore, four models of perceptual contrast were tested in simulations using auditorily-based spectra produced by a model which incorporates two levels of processing, (1) narrow-band auditory filtering (R. D. Patterson, *J. Acoust. Soc. Am.* 1976, **59**, 640) and (2) wide-band integration (L. A. Chistovich, *J. Acoust. Soc. Am.* 1985, **77**, 789). The experiment's results could be approximated by either of two models: an auditory figure/ground model, and a talker contrast model. A second experiment distinguished between these two models. The auditory figure/ground model predicts that in a cross-series anchoring experiment (in which tokens with high F_0 are used to anchor the low F_0 continuum and tokens with low F_0 are used to anchor the high F_0 continuum) the boundary of the vowel identification function will be shifted toward the vowel quality of the anchoring stimulus. The talker contrast model predicts that the vowel quality of the anchoring stimulus is less important than its F_0 and that the phoneme boundary will be shifted in the same direction regardless of the vowel quality of the anchoring stimulus. The results of the experiment quite unambiguously supported the predictions of the talker contrast model.

1. Introduction

Typically vowel normalization and vowel contrast have been studied separately. Researchers interested in vowel contrast have generally not considered vowel normalization processes (although see Fox, 1985), and researchers who have investigated vowel normalization have not considered the possible role of perceptual

0095-4470/90/020229 + 26 $03.00/0

contrast in the process of vowel normalization. This is especially interesting in view of the fact that fundamental frequency (F_0) normalization experiments involve intermixing tokens with different F_0. The research reported here is an attempt to gain a better understanding of vowel normalization (and, to some extent, vowel contrast) by determining the ways in which perceptual contrast and vowel normalization interact. The research is motivated by the assumption that we may be able to come to a better understanding of both of these important perceptual processes by considering the ways in which they interact.

1.1. *Vowel normalization*

Vowel normalization is a hypothetical perceptual process in which interspeaker vowel variability is reduced in order that perceptual vowel identification may then be performed by reference to relative vowel quality rather than to the absolute values of the acoustic parameters of vowels. It is well documented that hearers are influenced by F_0 when they make judgements about vowel quality (Miller, 1953; Fujisaki & Kawashima, 1968; Slawson, 1968; Ainsworth, 1975; Traunmüller, 1981). The general pattern of results reported in these studies is that the vowel formants must increase as F_0 is increased in order to maintain the same vowel quality. Because there is a correlation between vocal tract size and F_0, it is also possible to describe the effect in terms of the perception of the size of the talker's vocal tract. In this case, we would say that hearers perceptually normalize vowel formants by reference to some index of vocal tract size, using F_0 as a cue for this variable. Many researchers have assumed a similar process of normalization (the utilization of "an internal model of the speaker", Summerfield, 1971) and so, for this reason, the effect has been called vocal tract normalization.

As a cue for perceived speaker identity, F_0, in this view, may be used by hearers to establish a speaker-dependent perceptual vowel space. In this way, the expected acoustic correlates of particular vowel qualities are adjusted on the basis of the perceived identity of the talker. I will call this type of hypothesis an *adjustment-to-talker* model of vowel normalization. One way to implement an adjustment-to-talker model of vowel normalization is to use F_0 in order to estimate speaker-dependent formant ranges in a range normalization process (concerning range normalization see Gerstman, 1968). The model of vowel normalization described by Bladon, Henton & Pickering (1984) is also an example of an adjustment-to-talker approach. This is not obvious, at first, because they describe the model as "auditory", but it is actually a two-stage model. They propose that vowel normalization is accomplished by shifting the auditory spectra of vowels produced by female speakers down by 1 Bark before comparing them to spectral templates based on vowels produced by men. The auditory stage in the model is in the calculation of "auditory" spectra (Bladon & Lindblom, 1981). The second stage (when spectra are shifted along the Bark scale) involves an adjustment-to-talker.

Traunmüller (1981) proposed an explanation of F_0 normalization which is an alternative to the adjustment-to-talker view of normalization in that it appeals only to properties of the auditory system. No information about speaker identity or vocal tract length need be specified in this model.[1] He proposed that the "centre of

[1] Other approaches to perceptual vowel normalization which avoid reference to the size of the vocal tract include those discussed by Sussman (1986), Sydral & Gopal (1986), Nearey (1978) and Miller

gravity" effect reported by Chistovich, Sheikin & Lublinskaja (1979) can account for vowel normalization data.

Chistovich *et al.* (1979) reported that vowel formants which are within about 3 Bark of each other seem to be integrated into a single perceptual "centre of gravity". In their experiments, subjects manipulated the frequency of a synthetic single-formant vowel until it matched, as closely as possible, the quality of a two-formant standard. They found that when the formants of the standard were within about 3 Bark of each other, subjects tended to adjust the frequency of the single-formant so that it fell at the amplitude weighted mean of the formants of the standard. However, when the separation between the formants of the standard was greater than 3 Bark, subjects tended to adjust the single formant so that it matched one of the formants of the standard. These data are consistent with the observations of Delattre, Liberman, Cooper & Gerstman (1952), who reported that acceptable back vowels, in which F_1 and F_2 are close in frequency, can by synthesized with only a single formant. However, for front vowels, in which F_2 and F_3 are near each other, the F_2 of two-formant synthetic vowels must be higher than the F_2 found in natural speech. The data reported in Johnson (1989) are also consistent with the "centre of gravity" effect. Johnson reported that a higher formant normalization effect (see Fujisaki & Kawashima, 1968) is found when F_2 and F_3 are within 3 Bark of each other, but not when F_2 and F_3 are separated by more than 3 Bark.[2]

Traunmüller's proposal is that the effect of increasing F_0 on vowel perception results from an increased participation of the lowest harmonic in the "centre of gravity" which corresponds to the perceived F_1 (F_1'). As F_0 increases the lowest harmonic enters more and more into the window of integration which includes the peak of F_1. This increased influence of the lowest harmonic results in a lower F_1' as F_0 increases. Of course, as F_0 increases past the actual F_1, F_1' follows F_0. In the second section of this paper, I will consider whether spectra generated by an auditory model have the properties suggested by Traunmüller.

1.2. *Vowel contrast*

One of the earliest findings in the study of speech perception was that vowel discrimination is better than would be predicted by vowel labeling (Fry, Abramson, Eimas & Liberman, 1962; Eimas, 1963). These authors also found that the label given to an ambiguous vowel token is a function of context. When an ambiguous token from a continuum from A to B is preceded by the A endpoint the ambiguous token is more likely to be labeled B, and in the context of the B endpoint the label will be shifted to A. This *vowel contrast* effect, coupled with better-than-predicted

(1989). This research and that of Gerstman (1968) and Labonov (1971) (see Disner, 1980 for a review and critique) is concerned with normalization algorithms, and not explicitly with perceptual processing. I have chosen to emphasize models of vowel normalization which are more directly concerned with perceptual processing, but this does not indicate a disregard for the algorithm approach. Rather, I wish to avoid some of the assumptions of this approach: in particular the assumption that vowel normalization is preceeded by formant extraction. Clearly, vowel normalization algorithms which correctly classify vowels may reflect perceptual processes which operate in ways similar to those encoded in the algorithms (see Traunmüller, 1981; Neary, 1978; Sussman, 1986).

[2] Although the criticisms of the use of front vowel continua (see the discussion of Experiment 2) also apply to the work in Johnson (1989).

discrimination, led the Haskins group to conclude that vowels are perceived continuously because small changes in vowel quality may be easily produced, while consonants are perceived more categorically because their production tends to be categorical. The work of Fujisaki & Kawashima (1969) and Pisoni (1971, 1973, 1975) turned attention to the role of auditory memory in speech perception. It was hypothesized that vowels, which have longer, more steady-state acoustic cues, leave a longer lasting trace in auditory memory, and so the auditory differences between vowels are more readily available for use in speech perception tasks. Thus, a dual-process view of vowel perception (which involves both an auditory stage and a phonetic stage) provides an explanation of vowel context effects. Simon & Studdert-Kennedy (1978) called this view of vowel contrast an auditory figure/ground approach.

Feature detector theory (Eimas, Cooper & Corbit, 1973; Eimas & Corbit, 1973; Cooper, 1974) provides a possible explanation of vowel contrast in terms of feature-detector fatigue. However, as Fox (1985) points out: "this hypothesis has rarely been mentioned in connection with vowel perception, presumably because the large number of vowel categories and the relatively noncategorical perception of the stimuli made explanations in terms of discrete dectors seem unattractive. Also, while feature-detector fatigue is a plausible mechanism for explaining selective adaptation effects, it cannot account for pairwise contrast where only a single contextual item is presented" (p. 1552).

Crowder (1981) proposed a model of vowel contrast in which the memory representations behave "in accordance with the laws of *recurrent lateral inhibition*" (p. 175). In this model, the auditory memory representation of a stimulus interacts with that of an earlier stimulus in a process which is analogous to lateral inhibition in peripheral sensory systems. Crowder suggests that the frequency components of representations of stimuli interact with each other in auditory memory in such a way that overlapping frequency components are mutually inhibiting, whereas unique components are relatively uninhibited and tend to dominate in the classification of stimuli. Thus, in Crowder's view of vowel contrast, the spectral differences between stimuli are enhanced and similarities are inhibited as a result of the nature of their representations in auditory memory and the hypothesized process of lateral inhibition. An important feature of Crowder's proposal is the notion of "channels" in auditory memory. According to Crowder, items produced by different talkers will occupy different channels of auditory memory, with degree of perceived talker difference determining the degree of channel discrepancy. He suggests that items which are on the same or similar channels will inhibit each other, while items on different channels will not. Therefore, it is not clear how relevant this model of vowel contrast is to situations in which vowels with different speaker qualities are presented.

Simon & Studdert–Kennedy (1978) and Fox (1985) also considered a response bias explanation (Parducci, 1965, 1975) of vowel contrast. In this account, the change in labeling behavior in an anchoring experiment is due to the increase in the number of stimuli which must receive one label. Parducci's range-frequency theory predicts that subjects will attempt to use category labels an equal number of times during an experimental session, and so will show a boundary shift when anchoring stimuli are presented with a test continuum, because the anchor stimuli are

consistently labeled with one of the available labels. This seems to explain the results of anchoring experiments where one stimulus is presented more often than another, but it does not account for evidence of vowel contrast in experiments where each of the stimuli is presented equally often (Fry *et al.*, 1962). The fact that there is a shift in identification in anchoring experiments even when subjects are made aware of the relative frequency of occurrence of each token, also suggests that the response bias explanation is not an adequate explanation of this effect (for other arguments see Fox, 1985, p. 1553).

Finally, in addition to vowel contrast effects, I will consider in this paper the possibility that the perceived identity of the speaker may undergo talker contrast effects. That is, when two (synthetic) voices are placed close to each other in time, the degree of perceived difference between the voices may be larger than when they are temporally separate. As will be shown below, this type of contrast is an important consideration for experiments, such as those reported here, in which both vowel quality and speaker quality are manipulated.

2. Experiment 1

Most previous studies of F_0 normalization have one methodological trait in common: stimuli at different F_0 levels are presented intermixed with each other. This presentation format corresponds to one of the conditions studied by Mullennix, Pisoni & Martin (1989). In those experiments, they found that when hearers were required to identify words in different levels of noise, word recognition performance was impaired by random variation of talker identity. Performance was better when all of the words presented for identification had been produced by the same talker, as compared to a condition in which the identity of the talker varied from trial to trial. Mullennix *et al.* also found reliable reaction time differences between single-talker and multiple-talker conditions in two naming experiments. Subjects could repeat aloud words in the single-talker condition about 50 ms faster than they could in the multiple-talker condition (averaged over lexical density and word frequency conditions in two experiments). The reaction time data were interpreted as indicating that hearers must adjust to the talker in the multiple-talker condition, while this adjustment is not required in the single-talker case. Mullennix *et al.*'s experiments are relevant to the study of vowel normalization because they indicate that hearers do not automatically "normalize" the speech that they hear, but rather that some exposure to a new talker is required in order to be able to identify words as quickly and accurately as possible.

Previous research on F_0 normalization has involved the presentation of synthetic speech tokens in what is essentially a multiple-talker condition, though this "multiple-talker" condition has usually been composed of only two levels of F_0 (Miller, 1953; Fujisaki & Kawashima, 1968; Slawson, 1968). The present experiment extends the traditional format by including a condition in which tokens are blocked by F_0. In the analog of Mullennix *et al.*'s single-talker condition (here called the single-F_0 condition), the tokens were blocked by F_0, thus the F_0 of the tokens was entirely predictable within blocks. In the analog of their multiple-talker condition (here called the mixed-F_0 condition), the tokens of the two F_0 continua were randomly intermixed with each other.

2.1. *Method*

2.1.1. *Subjects*

Twenty-four undergraduate students at Indiana University participated in the experiment (18 female, 6 male). All were native speakers of American English who had never experienced any speech or hearing disorders. They received partial course credit in an introductory psychology course for their participation.

2.1.2. *Materials*

The stimuli used in this experiment were synthetic CVC syllables in a vowel continuum from [hɷd] to [hʌd]. Two continua were synthesized using a cascade-parallel formant synthesizer (Klatt, 1980), one continuum with a steady-state F_0 of 120 Hz, the other with steady-state F_0 of 240 Hz. The formant values of the synthetic vowels are shown in Table I and in Fig. 1. These formant values had been used in previous studies of vowel F_0 normalization (Johnson, 1989, in press). The syllables were 285 ms in duration. The aspiration noise of the /h/ was 95 ms long (with F_1 and F_2 slightly higher than the F_1 and F_2 of the vowel as naturally occurs as a result of tracheal coupling). The steady-state vowel portion of the stimuli was 160 ms long. Bandwidths of F_1, F_2, and F_3 were 110, 75, and 110 Hz, respectively. F_4 and F_5 were 3500 Hz and 4200 Hz, both with a bandwidth of 300 Hz, and were steady-state throughout the syllables. The final transitions into /d/ were 30 ms long and ended at 300, 1700, and 2516 Hz for F_1, F_2, and F_3, respectively. The F_3 transition dipped to 2116 over the first 15 ms of the transition and then rose to 2516. The peak amplitudes of the tokens were equated (to avoid possible effects of amplitude variation on reaction time).

2.1.3. *Procedure*

Stimuli from the two vowel continua were presented over TDH-39 headphones at listening level of 80 dB using two types of presentation. In the single-F_0 condition, the tokens were blocked by F_0. In the mixed-F_0 condition, the tokens from the two continua were randomly intermixed with each other. Each stimulus was presented ten times in each of these two conditions. Subjects were randomly divided into two groups. One group of subjects responded first to the items in the single-F_0 condition and then to the same items in the mixed-F_0 condition. The other group first heard the items in the mixed-F_0 presentation type and then responded to them again in the single-F_0 condition. The two groups will be called the single-first and mixed-first groups, respectively. In the single-F_0 condition, the order of presentation of F_0 level was counter-balanced across subjects. So, presentation type and F_0 level were

TABLE I. Formant values of the test tokens used in the listening experiment

	Token						
	1	2	3	4	5	6	7
F_1	474	491	509	526	543	561	578
F_2	1111	1124	1137	1150	1163	1176	1189
F_3	2416	2424	2432	2440	2448	2456	2464

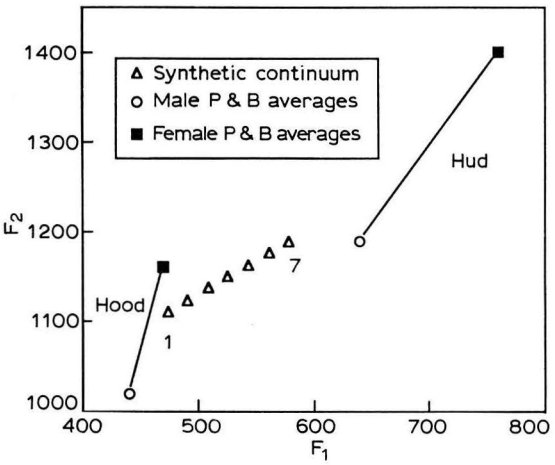

Figure 1. Tokens in the "hood"–"hud" continuum, and the Peterson & Barney (1952) average values for midwestern male and female speakers of English.

treated as within-subject variables while order of presentation was treated as a between-subjects variable.

A video monitor was mounted at approximately eye level for each subject.The words "hood" and "hud" were presented on the monitor at the left and right of the screen. Subjects used these labels as an indication of which button to press in a forced-choice identification task. For half of each subject's responses in each condition, "hood" was the right-hand response and "hud" was the left-hand response. For the other half of the trials, the right-hand response was "hud" and the left-hand response was "hood". Button to response associations were switched at intervals of 70 trials, so within a block of 70 trials the association was constant.

Each token in the two continua was presented 10 times in each of the two types of presentation. The number of presentations per subject was 280 (7 tokens × 2 F_0 levels × 2 presentation types × 10 presentations). Both identification and reaction time data were collected online by a PDP 11/34 mini-computer at the Speech Research Laboratory at Indiana University. Subjects participated in the experiment in groups of up to six at a time.

2.2. *Results*

2.2.1. *Identification data*

The identification data are shown in Fig. 2. The two presentation conditions are shown in separate graphs, the responses of the mixed-first group are plotted with open symbols, while the responses of the single-first group are plotted with filled symbols. Responses to the low F_0 continuum are plotted with circles and the responses to the high F_0 continuum are plotted with squares. The identification data were analyzed in a four-way repeated measures ANOVA with factors Presentation Type (mixed-F_0 *vs.* single-F_0), F_0 Level (120 Hz *vs.* 240 Hz), Group (mixed-first *vs.* single-first) and Token (1–7).

The only significant main effects in the analysis were for F_0 Level [$F(1, 22) =$

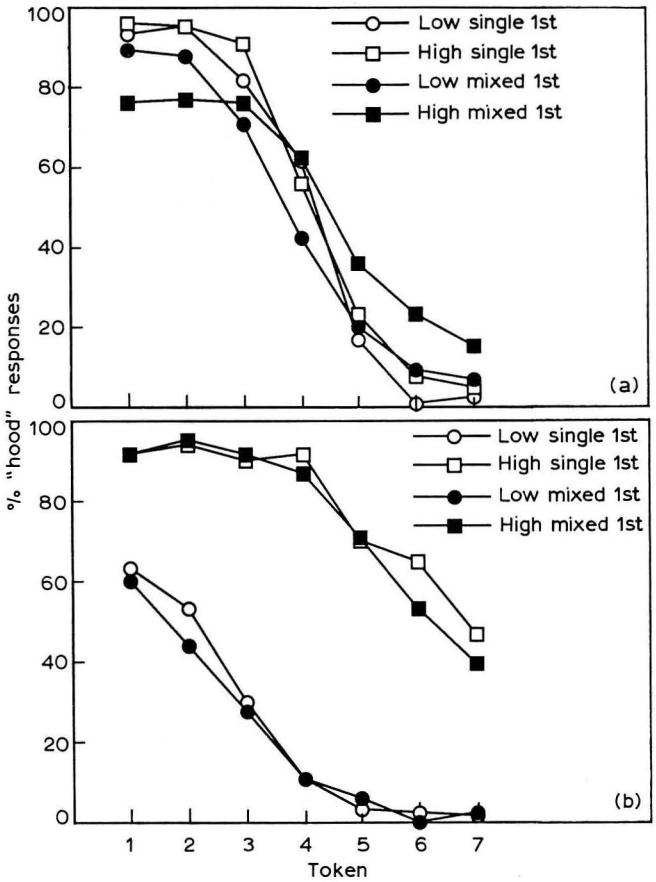

Figure 2. Identification data as a function of token number, presentation type, subject group and F_0 level. (a) Single-F_0 condition; (b) mixed-F_0 condition.

91.76, $p < 0.0001$] and Token [$F(6, 132) = 154.14$, $p < 0.0001$]. The low F_0 continuum was identified as "hood" 35.5% of the time while 64.9% of the high F_0 items were labeled "hood". The interaction of Presentation Type and F_0 Level was significant [$F(1, 22) = 123.07$, $p < 0.0001$]. This interaction can be observed in Fig. 2 as the different relationships between circles and squares in the top as opposed to the bottom panel. Table II shows the average percent "hood" responses to each continuum (low and high F_0). When items which differed in F_0 were presented intermixed with each other there was a large difference between the identification functions as a function of token F_0, whereas when the items were presented in separate blocks there was no effect of F_0 identification behavior. The three-way interaction of Presentation Type, F_0 Level and Token was also significant [$F(6, 132) = 4.82$, $p < 0.001$]. Examination of the functions in Fig. 2 indicates that this interaction occurred because the effect of F_0 in the mixed-F_0 condition was to shift the boundary between "hood" and "hud" and not to change the probability of "hood" responses globally across the continuum.

The interaction between Presentation Type and Token was also significant [$F(6, 132) = 32.61$, $p < 0.0001$]. The average identification function in the mixed-F_0

TABLE II. The interaction of Presentation Type and F_0 Level. The data in this table are percent "hood" identifications as a function of presentation type and F_0 level averaged across subjects and tokens

	Low F_0	High F_0
Single-F_0	48.4	52.7
Mixed-F_0	22.6	76.9

condition was flatter than was the average identification function in the single-F_0 condition. Also, there was an interaction between F_0 Level and Token [$F(6, 132) = 15.33$, $p < 0.0001$]. The effect of F_0 (averaged over groups and presentation conditions) was a boundary shift and not a global change in probability of "hood" response.

Finally, there were two significant interactions which involved group differences. As is clear in Fig. 2(b), the two groups of subjects (mixed-first and single-first) had virtually identical response functions in the mixed-F_0 condition, while in the single-F_0 condition [Fig. 2(a)], the response functions for the mixed-first group were somewhat flatter than those of the single-first group. This difference was reflected in the Presentation Type by Token by Group interaction [$F(6, 132) = 4.25$, $p > 0.001$]. Also, in Fig. 2(a) (the single-F_0 condition), it appears that this group difference was larger for the high F_0 level than for the low F_0 level (i.e. the function for the high F_0 mixed-first continuum is flatter than the function for the low F_0 mixed-first continuum). This was reflected in the four way interaction of Presentation Type, Group, F_0, Level and Token [$F(6, 132) = 3.02$, $p < 0.01$]. It is not clear why the subjects in the mixed-first group would show less categorical identification of the continua in the single-F_0 condition than subjects in the single-first group.

2.2.2. *Reaction time data*

Average reaction times, measured from stimulus onset and averaged across both response and token, were analyzed in a repeated measures ANOVA with factors F_0 Level, Presentation Type and Group. The only statistically significant effect in this analysis was the main effect for Presentation Type [$F(1, 22) = 5.39$, $p < 0.05$]. Average reaction time in the mixed-F_0 condition was 697 ms and in the single-F_0 condition was 647 ms. The main effect for F_0 Level approached significance [$F(1, 22) = 3.76$, $p = 0.0653$]. The trend was for items with low F_0 to be identified more quickly than the items with high F_0. This effect may relate to the relative naturalness of the different levels of F_0, since the tokens sounded more natural at lower F_0 levels.

An additional analysis of the reaction time data from the mixed-F_0 condition assessed the effect of changing F_0 from token to token. Classifying the reaction time data by the F_0 of the item being identified and by the F_0 of the immediately preceding item results in four classes of reaction times; low F_0 items which were immediately preceded by a low F_0 item, low F_0 items which were immediately preceded by a high F_0 item, high F_0 items which were immediately preceded by a low F_0 item and high F_0 items which were immediately preced by a high F_0 item.

Analysis of these data in a three-way repeated measures ANOVA with factors: Token-F_0 (high or low), Context-F_0 (high or low) and Group (mixed-first or single-first) revealed one reliable main effect, namely Token-F_0 [$F(1, 22) = 9.96$, $p < 0.01$]. This effect is consistent with the marginal effect for F_0 Level found in the overall analysis. There was also a significant interaction between the Token-F_0 and Context-F_0 factors [$F(1, 22) = 5.98$, $p < 0.05$]. When there was a change of F_0 from one token to the next, subjects were slower to identify the token than when F_0 did not change from one token to the next.

2.3. *Discussion*

The difference in reaction time between the mixed-F_0 and single-F_0 conditions which was observed in this experiment (50 ms) is comparable to the reaction time difference observed by Mullennix *et al.* (1989) between their single-talker and multiple-talker conditions. A reaction time difference for blocked *vs.* mixed voices was also reported by Summerfield & Haggard (1975). They found a reaction time difference which could be attributed to a normalization process, as opposed to a general effect of divided selectional attention, when F_0 and F_3 varied together, but not when F_0 varied alone. Although it is possible that the reaction time difference found here reflects a normalization process (F_0 variation was much greater in this study than in that of Summerfield & Haggard), the proper control conditions were not included in this study, and so it would be premature to claim that the reaction time difference is evidence for a special normalization process.

The identification data indicate that when tokens from vowel continua with different F_0 are presented randomly intermixed with each other there is an effect of F_0 upon vowel identification; however, when tokens are presented blocked by F_0 the effect of F_0 is severely diminished (if present at all). This pattern of results suggests the operation of a contrast effect. In the sections that follow I will attempt to determine what type of contrast effect can account for these data.

3. Model studies of vowel normalization and vowel contrast

This section is organized into three parts. Section 3.1 is description of a model of the auditory representation of vowels. In Section 3.2, spectral representations generated by this model are used to investigate the effect of F_0 on the auditory representation of vowels. In particular, the predictions of Traunmüller's (1981) hypothesis concerning vowel normalization are tested. Section 3.3 reports the results of four simulations of the mixed-F_0 condition of Experiment 1. The simulations implement different approaches to perceptual contrast.

3.1. *An auditory model*

The auditory model described here incorporates some of the frequency and amplitude nonlinearities found in psychophysical studies of auditory processing, and a spectral integration stage which simulates Chistovich *et al.*'s (1979) "centre of gravity" hypothesis. The work of Schroeder, Atal & Hall (1979) and Bladon and Lindblom (1981) formed the foundation of the approach I adopted here. In particular, the use of spectral integration is an important hypothesis concerning the way in which hearers estimate the broad features of the spectral envelopes of speech

sounds as those broad features relate to the vocal tract transfer function. It is important to keep in mind, however, that this model is only a *rough* implementation of some hypotheses concerning the human auditory treatment of speech sounds.

The first stage involves the calculation of the magnitude spectrum of a Hamming window of speech samples. In the second stage, the magnitude spectrum is conditioned by a bank of filters. Following Patterson (1976), the equivalent rectangular bandwidth (BW_{ER}) of the filters is given by (1), and the auditory filter shape is given by (2). In these equations f_0 refers to the center frequency of the filter in Hertz.

$$10 \log_{10} BW_{ER} = 8.3 \log f_0 - 2.3 \tag{1}$$

$$|H(\Delta f/f_0)^2| = \exp[-\pi(\Delta f/f_0 BW_{ER})^2] \tag{2}$$

These filter shapes are Gaussian approximations to the filter shapes determined by psychoacoustic masking studies (see Moore & Glasberg, 1983). Filter functions were calculated at intervals of 0.2 Bark for the range 0.2 to 19 Bark (18 to 4884 Hz). The output of each filter (A_j) is determined by (3), where n is the number of terms in the filter, W_{ij} is the ith term of the jth filter, and S_i is the spectral magnitude at the frequency corresponding to the ith term of the jth filter. The sum of the products of the magnitude spectrum and the filter weights is normalized by the total of the filter weights because the number of terms in each filter is a function of center frequency.[3]

$$A_j = \sum_{i=0}^{n} S_i W_{ij} \Big/ \sum_{i=0}^{n} W_{ij} \tag{3}$$

Next, a stored equal-loudness contour (Fig. 3) is applied to the spectrum. Each frequency location in the filtered spectrum is attenuated by the amount indicated in the equal loudness contour, with low frequency components being attenuated more than others.

Figure 4(a) shows the Fourier transform of a synthetic [ɷ] with an F0 of 120 Hz. Figure 4(b) shows the same spectrum after passing it through the filter bank and applying the equal-loudness contour. Note that in the filtered spectrum, all but the

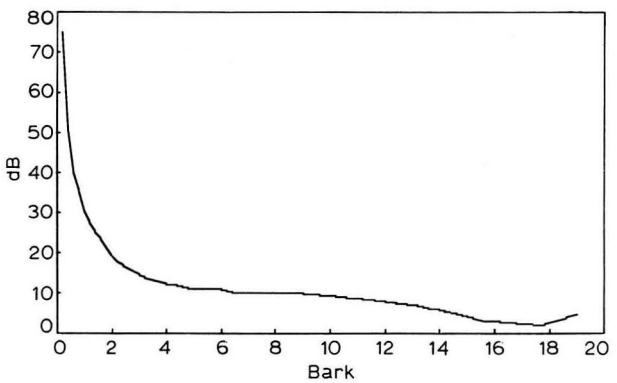

Figure 3. Equal loudness contour, derived from Fletcher & Munson (1933). Ordinate indicates degree of attenuation at each frequency (abscissa).

[3] This is necessary because the bandwidths of the filters increases as a function of the center frequency. The samples in the Fourier transform are linearly spaced in frequency. Therefore, as the bandwidths of the filters increases, the number of samples under the filter window increases.

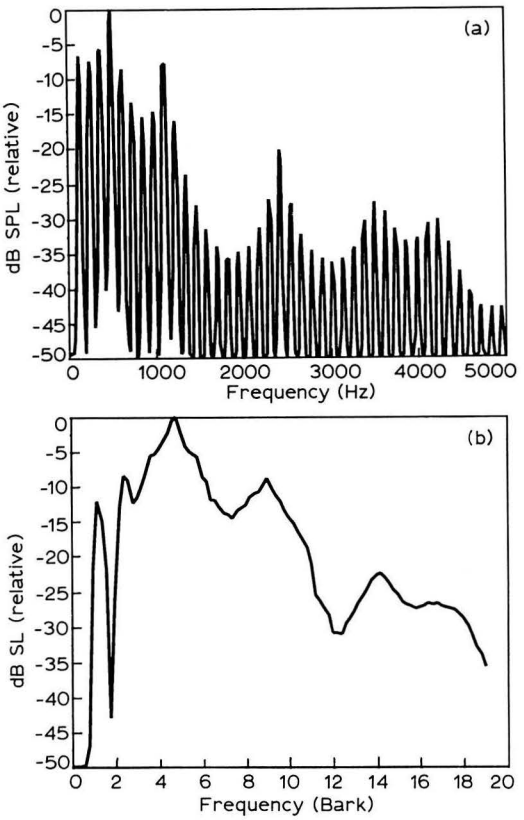

Figure 4. Spectra of the vowel [ɷ] of "hood". (a) The Fourier transform. (b) Filtered, loudness-equalized spectrum.

two lowest harmonics are smeared together. This is a result of the increasing bandwidths of the auditory filters; frequency resolution decreases as frequency increases. Note also that the regions for F_1 and F_2 are expanded (as compared with the Fourier transform) and that the higher frequency regions are compressed. Both of these observations are also true of displays of auditory nerve responses to vowel sounds (Sachs & Young, 1979).[4] The model, at this stage of processing, captures some basic properties of peripheral auditory processing.

The third stage of the model involves sliding a window of integration across the spectrum. This stage is an implementation of Chistovich *et al.*'s (1979) hypothesis that, in vowel perception, spectral components over a fairly large spectral range are integrated into a single "centre of gravity". There is no evidence of spectral integration of this sort in neural responses in the auditory pathways of animals (even in the tonotopically organized regions of the cortex, see Pickles, 1988); therefore, if there is a stage of spectral integration in human auditory processing, it is most likely

[4] Young & Sachs (1979) emphasized also the fact that firing rate saturates, and consequently that spectral specificity is lost at moderate amplitudes. They suggest that temporal measures such as localized phase locking must also be involved in auditory frequency resolution. Thus, the similarities between the filtered, loudness-equalized spectrum (based on psychophysical studies of hearing) and published displays of mean firing rate or Average Localized Synchronized Rate (ALSR), for that matter, are at best merely suggestive.

a speech specific, central auditory process (see Chistovich, 1985; Traunmüller, 1982).

In this model, the Riemann sum (4) over a portion of the spectrum served as an approximation to the definite integral for that spectral region. In (4), Δx_k was 0.2 Bark for all k (the interval between samples in the filtered spectrum) and $f(t)$ was the filtered, loudness-equalized spectrum [Fig. 4(b)]. The program calculated Riemann sums over successive windows in the filtered spectrum to produce a power density spectrum. The y dimension for each frequency bin of the power density spectrum was the Riemann sum over a 2.2 Bark window centered on that frequency, and separate sums were calculated at intervals of 0.2 Bark. The sums are expressed in decibels normalized to the RMS amplitude of the original waveform in order to preserve relative amplitude differences across speech samples. Experimentation with previous versions of the model indicated that integration over a window of 2.2 Bark resulted in a single spectral peak between F_1 and F_2 when they were within three Bark of each other. Wider integration windows resulted in the merger of formants which were separated by more than 3 Bark.

$$\sum_{k=1}^{n} f(t_k)\,\Delta x_k \tag{4}$$

Because the spectra generated by this model are based on a *rough* attempt to implement some hypotheses concerning the auditory processing of speech, I will refer to spectra produced by the model as "auditorily-based" (AB) spectra. The following section examines the effect of F_0 in AB spectra, and Section 3.3 reports the results of some simulations of mixed condition of Experiment 1.

3.2. *The effect of F_0 in simulated auditory spectra*

If F_0 normalization can be attributed to the "centre of gravity" effect, as suggested by Traunmüller (1981), the model described above should produce spectra in which, as F_0 increases, the spectra of "hud" tokens become more like "hood", and as F_0 decreases, the spectra of "hood" tokens become more like "hud".

Consider first the AB spectra of the "hood-hud" continuum with low F_0 [Fig. 5(a)]. There appear to be two primary differences between "hood" and "hud". First, the frequency locations of spectral peaks are lower for "hood" and, second, the amplitude of the spectrum above about 6 Bark is higher for "hud". Figure 5(b) shows the AB spectra of the continuum with high F_0. In these spectra, the first peak occurs at the same frequency throughout the continuum. The difference between "hood" and "hud" for these tokens is mainly in the amplitude of the components around 8 to 10 Bark and the amplitude of the first peak itself ("hood" has the higher amplitude first peak).

Figure 6 illustrates the effect of F_0 on the AB spectra of these vowels. This figure compares the AB spectra of vowels which have identical formant values and different F_0 values. The top panel shows the effect of F_0 on the spectra of the "hud" endpoint of the continuum. As F_0 increases, the frequency of the first spectral peak decreases. This is exactly what an auditory model of F_0 normalization would predict. The bottom panel of the figure shows the AB spectra of the "hood" endpoint of the continuum. Here there is no correlation between F_0 level and the location of the first

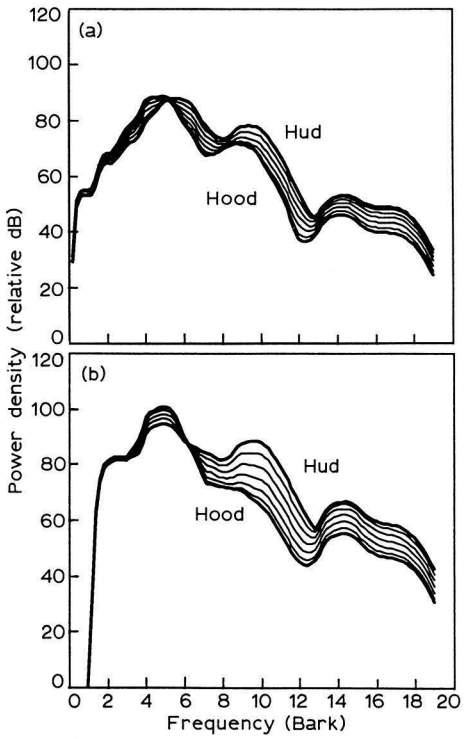

Figure 5. AB spectra of the vowels in the low and high F_0 "hood-hud" continua which were used in the Experiment 1. (a) Low F_0 continuum. (b) High F_0 continuum.

spectral peak. This finding is especially troublesome for the auditory approach because, with its lower F_1, "hood" should be more sensitive to F_0 than "hud".

These data do not lead to the conclusion that as F_0 increases AB spectra of "hud" become more "hood"-like, or that as F_0 decreases AB spectra "hood" become more "hud"-like (Traunmüller, 1981), but rather that, as F_0 increases, the AB spectrum is more and more determined by the harmonics of the fundamental. This is true for AB spectra as well as for simple Fourier transforms. Consider the frequency locations of the harmonics relative to the formant peaks. A vowel with a fundamental frequency of 120 Hz will have harmonics at 1.29, 2.53, 3.7, 4.79, 5.78 and 6.83 Bark. The F_1 of the synthetic "hood" was 4.74 Bark and of "hud" 5.6 Bark. When F_0 is low, the harmonics are closely spaced and, thus, there are harmonics near the F_1 of both "hood" and "hud". When the fundamental frequency is 180 Hz, there are harmonics at 1.9, 3.7, 5.3 and 6.69 Bark. So, the third harmonic is close to the F_1, of "hud", but F_1 falls between harmonics in "hood". Note the broad, flat peak in the AB spectrum of "hood" synthesized at 180 Hz [Fig. 6(b)]. When the fundamental frequency is 240 Hz, there are harmonics at 2.53, 4.79 and 6.83 Bark. In this case, the only harmonic in the region of F_1 (for these two vowels) is the second harmonic. In the AB spectrum of "hood" the second harmonic and F_1 are almost identical [note the amplitude of the peak in Fig. 6(b)]. While in the AB spectrum of "hud", the second harmonic is still closer to F_1 than are the other harmonics, but the two are not aligned (again the amplitude of the peak seems to

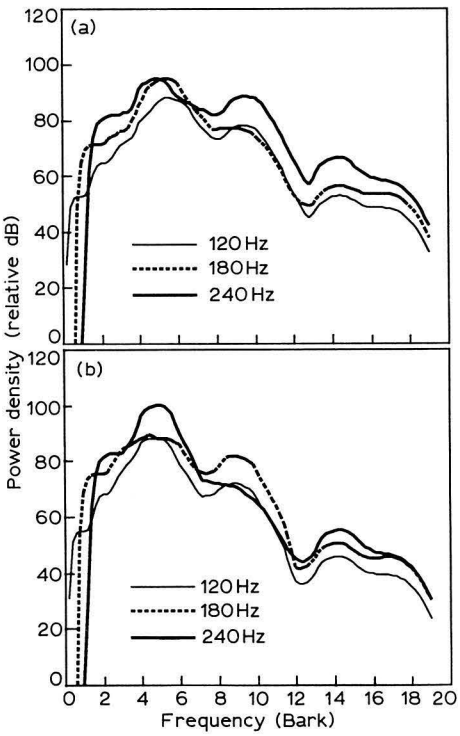

Figure 6. (a) AB spectra of "hud" with F_0 of 120, 180 and 240 Hz. (b) AB spectra of "hood" with F_0 of 120, 180 and 240 Hz.

reflect this [Fig. 6(a) and Fig. 5]. Thus, in the case of vowels with high F_0, the shape of the AB spectrum (like other spectral representations) is determined by both the harmonics of the fundamental and by the vowel formants. This investigation of hypothetical perceptual representations of vowel spectra offers no support for Traunmüller's (1981) hypothesis that F_0 normalization is the result of the auditory integration of F_1 and F_0 in a single "centre of gravity". Of course, the validity of this conclusion is dependent upon the validity of the model. The fact that the model produces spectra with a single spectral peak between F_1 and F_2 when they are within 3 Bark of each other suggests that it does capture Chistovich *et al.*'s (1979) proposal, and, thus, is appropriate for testing Traunmüller's hypothesis.

3.3. *Model studies of contrast in vowel perception*

The studies reported in this section evaluate the success of four different models of perceptual contrast in accounting for the data from the mixed condition in Experiment 1 (Fig. 2(b)]. AB spectra of the vowel tokens used in Experiment 1 above (see Fig. 5) served as the input representations in the models of contrast. The first is an implementation of Crowder's (1981) model. The second and third models implement an auditory figure/ground model of vowel contrast. AB spectra in these models were normalized before being subjected to contextual influence (following a suggestion by Fox, 1985, p. 1557). The second model includes an implementation of

Bladon *et al.*'s (1984) spectral shifting approach to vowel normalization and the third model includes an implementation of Gerstman's (1968) range normalization approach. The fourth model incorporates a talker contrast model implied by Mullennix *et al.*'s (1989) suggestion that hearers adjust to different talkers.

Luce's (1959) choice rule forms the basis of the decision component in each of these models. This rule is defined by (5). In this formula, S_{ij} refers to the similarity of token i to category j and b_j refers to response bias for category j. (b_j was held fixed at 0.5 in the first models and was allowed to vary in the last model.) S_{ij} was calculated by taking the inverse of spectral distance between the test spectrum and a spectral template. Distance was calculated by (6) after removing the DC offset for differences in overall amplitude. In this formula, S_i is the AB spectrum of token i and S_j is the AB spectrum of template j. The interval *ab* excludes the lowest two Bark and the highest two Bark because incomplete integration windows spanned these edge samples. Equation (6) is similar to the spectral distance metrics used by Plomp (1976, p. 95) and Bladon & Lindblom (1981). However, where the Euclidian or the city block measures of spectral distance was used in their metrics, the mean squared distance was used here. When the inverse of this measure of distance was used as S_{ij} in (5), the resulting response functions were very similar to those found in the blocked condition of Experiment 1 [Fig. 2(a)]. This similarity is an important starting point for the simulations of the mixed condition [Fig. 2(b)].

$$P(R_{ji} \mid T_i) = b_j S_{ij} \Big/ \sum_{j=1}^{2} b_j S_{ij} \tag{5}$$

$$D_{ij} = \left(\sum_{x=a}^{b} (S_i(x) - S_j(x))^2 \right) \Big/ (b - a) \tag{6}$$

3.3.1. *Crowder's model of contrast*

In the implementation of Crowder's model of contrast, AB spectra from the high and low "hood"–"hud" continua were classified based on a comparison with stored templates for "hood" and "hud". The average of the AB spectra of the high- and low F_0 endpoints of the continua served as templates in this model. The model classified each vowel in the context of every other vowel (both within and across continua), after the context spectrum had been attenuated by a certain proportion (the decay parameter), and then subtracted from the test spectrum. If the context spectrum has very little energy at a particular frequency, then the test spectrum will remain relatively unchanged at that frequency. However, if the two spectra have peaks in about the same location in frequency, then the peak of the test spectrum will be reduced (to an extent determined by the decay parameter) as a result of context. The value of the decay parameter which provided the best fit to the data of Experiment 1 [Fig. 2(b)] was estimated by the method of least squared error. No value of the decay parameter provided a very close fit to the data. The RMS error of the best fit obtained was 35.4 (Table III).

This simulation indicates that a model of vowel contrast along the lines of that proposed by Crowder (1981) does not account for the contrast effect found in the Experiment 1. This may be an indication that different talkers should be viewed as occupying different "channels" in auditory memory and that we should not expect recurrent lateral inhibition to play a role in contrast when two different voices are involved.

TABLE III. RMS data and parameter estimates for the model studies. In all model fits, the RMS error is in the same units as the data—i.e., percent "hood" responses

Model	RMS error	Parameter values	
Lateral inhibition	35.4	decay = 0.998	
Figure/ground contrast			
Spectral shifting	19.8	shift = 0.8	
Range normalization	19.8		
Figure/ground with lateral inhibition		shift = 0.8	decay = 0.3
Spectral shifting	10.3	decay = 0.3	
Range normalization	10.3		
Talker contrast	8.9	bias = 0.75	

3.3.2. *Auditory figure/ground contrast*

In the auditory figure/ground models of vowel contrast, the probability of a "hood" response for context items influenced the probability of a "hood" response on the current item. If the immediately preceding item was very much like "hood", then an ambiguous item will be more likely to be identified as "hud" than if the context item was a good example of "hud". In this model of vowel contrast (7), the adjusted probability of a "hood" response (Π_n(hood)) was equal to the base probability of the current item (defined by (5)) multiplied by the ratio of the base probability of the current item and the base probability of the immediately preceding context item, with resulting probability values truncated to the range 0 to 1. Each token served as a context for every other token in computing the average probability of a "hood" response.

$$\Pi_n(\text{hood}) = P_n(\text{hood}) \times (P_n(\text{hood})/P_{n-1}(\text{hood})) \tag{7}$$

Prior to the calculation of context effects, the AB spectra were normalized using an implementation of one of two different approaches to vowel normalization. One model used an implementation of Bladon *et al.*'s spectral shifting model of normalization. In this approach to normalization, AB spectra of vowels with high F_0 were shifted down on the Bark scale and then compared with AB spectral templates appropriate for a male speaker. AB spectra of steady-state tokens synthesized using the Peterson & Barney (1952) average formant and F_0 values for male "hood" and "hud" served as vowel templates in this model. The dialect of the speakers in Peterson and Barney's study was similar to that of the subjects in Experiment 1, so the use of these values is appropriate. The second implementation of the auditory figure/ground model used a form of range normalization (Gerstman, 1968). In this implementation of vowel normalization, the choice of templates depended on F_0. If F_0 was low, the Peterson and Barney average male templates were used. If F_0 was high, the templates were derived from the Peterson and Barney average vowel formants and F_0 for females.

In the spectral shifting model, the degree of shift was a free parameter. The degree of spectral shift which was estimated by the least squared error method was 0.8 Bark. This corresponds quite closely to the value used by Bladon *et al.* (1984). There were no free parameters in the range normalization model. Although these two approaches to vowel normalization classify the continua in quite different ways in the absence of any contrast effect (the spectral shifting model classifies both

continua more consistently), they provide virtually identical results when used in the vowel contrast model. RMS error of both the spectral shifting model and the range normalization model was 19.8. Both of these models provide a better fit to the data than does the lateral inhibition model, but the predictions are still pretty rough. However, when the lateral inhibition context effect is included in these models, the degree of fit improves considerably. RMS error was 10.3 for both the spectrum shifting and range normalization models. Predicted identification functions and the data obtained in the mixed condition of Experimental 1 are shown in Fig. 7(a).

Lateral inhibition affects the spectral representation of the stimulus before it is compared with a spectral template, and figure/ground contrast affects the decision rule used to classify the stimulus after spectral similarity has been calculated. It is, therefore, reasonable to expect that both lateral inhibition in auditory memory and figure/ground contrast could be involved in vowel contrast effects since they occur at different stages of processing. Note, however, that this conclusion, unlike the one reached in the previous section, suggests that different voices are interacting in auditory memory (perhaps across "channels").

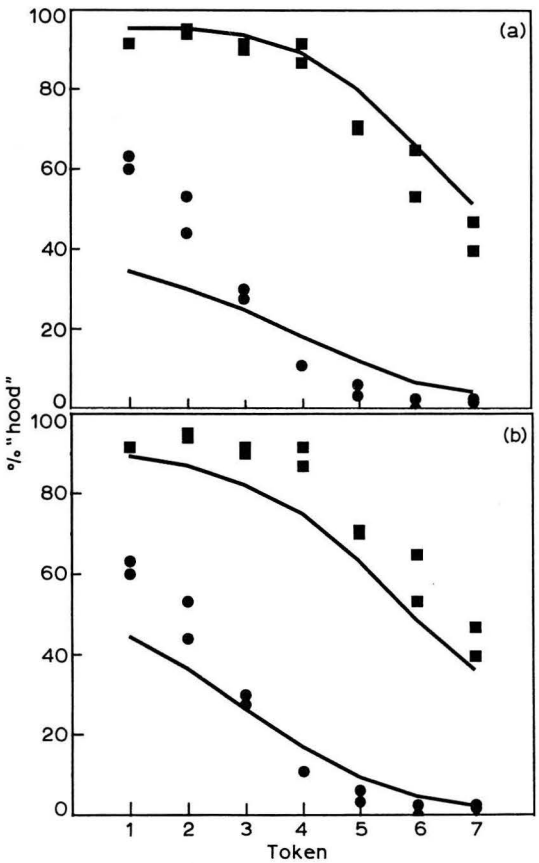

Figure 7. Results of model simulations of the mixed condition of Experiment 1 using (a) a post-normalization vowel quality contrast model, and (b) a talker contrast model. Model predictions are the solid lines and the data from Experiment 1 [see Fig. 2(b)] are plotted with filled squares (high F_0) and circles (low F_0).

3.3.3. *Talker contrast*

The final model implements a model of talker contrast. If hearers adjust their expectations for vowel quality relative to the perceived identity of the talker, as suggested indirectly by Mullennix *et al.* (1989) and more directly by Johnson (in press), these expectations could result in a contrast effect which depends on perceived speaker characteristics rather than on vowel quality. In other words, the contrast effect observed in Experiment 1 can be considered a talker contrast effect rather than a vowel contrast effect.

In the implementation of this view, the response bias factor (b_j) in the decision rule (5) varied as a function of the F_0 of the item. Low F_0 items had to be quite similar to "hood" in order to be classified as "hood", while high F_0 items had to be quite similar to "hud" in order to receive that label. This reflects the hypothesis of an adjustment-to-talker model of vowel normalization that the criteria for vowel classification are a function of the perceived identity of the talker. There was a single free parameter in the model. This parameter (the bias parameter) functioned as b_{hood} when F_0 was high and b_{hud} when F_0 was low. Bias toward the other category in each case was equal to 1–(bias). The spectra derived from the Peterson & Barney (1952) average formant values for male speakers were used as templates in this model.

Response probabilities predicted by the talker contrast model are shown in Figure 7(b). As the Figure shows, this model of contrast is also quite accurate in predicting the data of Experiment 1 (the RMS error was 8.9).

3.4. *Summary*

The findings of the model studies are summarized below:

(1) The hypothesis that vowel normalization is a consequence of auditory processing was not supported. In spectra generated by a model which incorporates two stages of filtering, as in other forms of frequency analysis, when F_0 increases, the shape of the spectrum becomes more and more dependent upon the harmonics of the fundamental frequency.

(2) Crowder's (1981) model of contrast cannot account for the results of Experiment 1. However, in combination with an auditory figure/ground contrast model it does help provide a close fit to the data. Unresolved is the question of whether vowels produced by different talkers should be seen as interacting in auditory memory.

(3) Range normalization (as implemented here) and normalization by means of spectral shifting (Bladon *et al.*, 1984) give identical fits to the data of Experiment 1 when they are used to provide the input to a figure/ground contrast mechanism.

(4) Two types of contrast provide good fits to the data—(1) auditory figure/ground contrast coupled with recurrent lateral inhibition and (2) talker contrast. The first involves contrast in vowel quality, the second in perceived talker identity.

4. Experiment 2

The two types of contrast which best account for the mixed-F_0 data of Experiment 1 make very different predictions for a cross-series anchoring experiment. If the

contrast effect observed in Experiment 1 occurred primarily at the level of vowel quality, we predict that the perceived vowel quality of the anchor token will be the dominant factor in cross-series anchoring. Conversely, if the contrast effect found in Experiment 1 was the result of a talker contrast process, we predict that the vowel quality of the anchor will be of less importance than the perceived identity of the talker, and thus, that the same direction of boundary shift will be produced by anchors of different vowel quality.

Experiment 2 is a test of these predictions. The stimuli which were used in Experiment 1 were presented in a cross-series anchoring experiment. Subjects heard the items of one of the two continua (low or high F_0) and then heard those same stimuli randomly intermixed with multiple occurrences of an anchor stimulus drawn from the other continuum.

4.1. *Method*

4.1.1. *Subjects*

Thirty-eight undergraduate students (10 male, 28 female) at Indiana University participated in the experiment for partial course credit in an introductory psychology course. All were native speakers of American English who had never experienced any speech or hearing disorders.

4.1.2. *Materials*

This experiment employed the same stimuli which had been used in Experiment 1.

4.1.3. *Procedure*

Subjects were randomly divided into four groups (two groups of 11 and two groups of 8). Each group was presented with a randomized list containing 20 repetitions of each of the tokens from either the high F_0 or low F_0 continuum in a control condition and then with those same tokens randomized with 60 occurrences of an anchor token from the other continuum. Thus, there were 140 trials in the control condition and 200 trials in the anchor condition. The first group of 11 subjects heard the low F_0 continuum in the control condition and the low F_0 continuum with 60 occurrences of token 1 from the high F_0 continuum in the anchor condition. Group two (11 subjects) also heard the low F_0 continuum, but with token 7 from the high F_0 continuum as an anchor. Groups three and four (8 subjects in each group) responded to the tokens of the high F_0 continuum with tokens 1 and 7 (respectively) of the low F_0 continuum as anchors. The equipment used to run the experiment was the same as that used in Experiment 1.

4.2. *Results*

The results of Experiment 2 are shown in Figs 8 and 9. Figure 8 shows the identification responses plotted by token number. Panel (a) shows the data for the low F_0 continuum for both the "hood" and "hud" anchor groups. The data presented in this panel were analyzed in a three-factor, repeated-measures analysis of variance. Factors were Condition (control vs. anchor), Anchor ("hood" *vs.* "hud"), and Token. There was (predictably) a main effect for Token [$F(6, 60) = 199.27$, $p < 0.001$]. More to the point, there was also a main effect for Condition [$F(1, 10) = 14.4$, $p < 0.01$]. In the control condition, the average percent "hood"

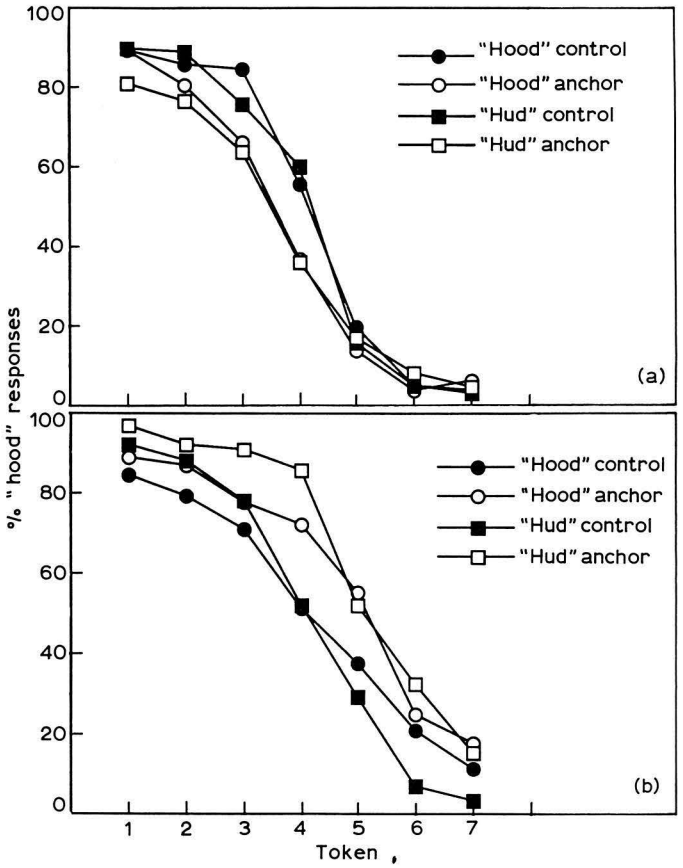

Figure 8. Results of Experiment 2. Identification functions for (a) the low F_0 "hood"–"hud" continuum, control conditions (solid points) and the anchor conditions (open points), and (b) the high F_0 "hood"–"hud" continuum.

response was 48.5%, while in the anchor condition this was reduced to 41.6%. The only other effect which reached significance was the Condition by Token interaction [$F(6, 60) = 8.52$, $p < 0.001$]. The effect of anchoring was to shift the phoneme boundary rather than to produce a global change in probability of a "hood" response.

Figure 8(b) shows identification functions for the high F_0 continuum. These data were also analyzed in an ANOVA with factors Condition, Anchor, and Token. The same three statistical effects were significant in this analysis. There were main effects for Token [$F(6, 42) = 116.2$, $p < 0.001$] and Condition [$F(1, 7) = 22.42$, $p < 0.01$], and the Condition by Token interaction was significant [$F(6, 42) = 3.91$, $p < 0.01$]. The Condition by Anchor interaction approached significance, but was not reliable [$F(1, 7) = 2.08$, $p = 0.19$]. This interaction is most visible in Fig. 8b as the difference between the "hood" and "hud" anchor conditions for tokens 3 and 4.

4.3. *Discussion*

These results quite clearly conform to the predictions of the talker contrast model; the direction of boundary shift did not depend on the vowel quality of the anchor.

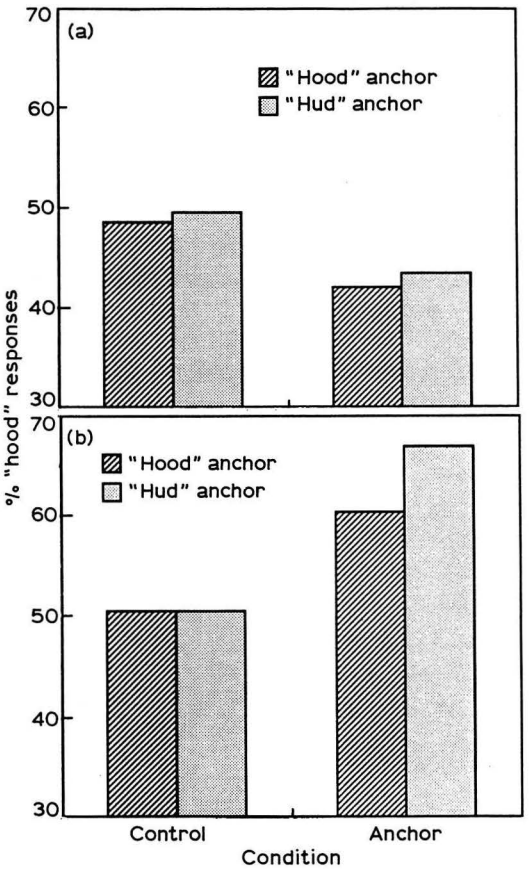

Figure 9. Results of Experiment 2 averaged over tokens. Percent "hood" responses to (a) the low F_0 "hood"–"hud" continuum, and (b) the high F_0 "hood"–"hud" continuum.

One concern should be addressed, though. The "anchor" tokens in this experiment were not uniformly identified as "hood" or "hud" in Experiment 1. In particular, the "hud" endpoint token was identified (in the mixed condition) as "hud" only slighly more than 50% of the time when it had high F_0. Also, the "hood" endpoint token of the low F_0 continuum (again in the mixed condition) was identified as "hood" only 60% of the time. Similarly, in the present experiment, when the anchor stimulus had high F_0, the "hood" anchor was identified as "hood" 88.7% of the time, while the "hud" endpoint anchor was identified as "hud" only 46.3% of the time. When the anchors had low F_0, the "hood" endpoint anchor was identified as "hood" 49.8% of the time and the "hud" endpoint anchor was identified as "hud" 97.4% of the time. Thus, these tokens do not satisfy the requirements of a model which relies on contrasting vowel quality.

However, not all of the subjects identified the anchor tokens in this manner. There were three subjects in the high F_0 "hud" anchor group who reliably (i.e. more than 75% of the time) identified the anchor as "hud". An analysis of variance of the data from these subjects also supported the talker contrast hypothesis. Factors were Token and Condition. There was a reliable interaction between these factors which

indicated a boundary shift [$F(6, 12) = 3.27$, $p < 0.05$]. The direction of this boundary shift was the same as the shift found in the overall analysis of variance. When the anchor had high F_0 and the continuum low F_0 the subjects tended to label ambiguous stimuli more as "hud" than they did in the control condition.

Of the subjects in the low F_0 "hood" anchor group, there were three who reliably labelled the anchor as "hood". An analysis of the data from these three subjects revealed a similar trend. In this analysis the Condition main effect approached significance [$F(1, 2) = 4.38$, $p = 0.17$]. The direction of this trend conformed with the overall analysis. When high F_0 items were identified in the context of low F_0 anchors (even anchors which were identified as "hood"), subjects tended to respond "hood" more frequently than they did in the control condition.

Note also that if the overall data were determined by vowel quality contrast, there should have been no boundary shift when ambiguous anchors were used. The fact that two stimuli which had quite ambiguous vowel qualities produced boundary shifts just as large (and in the same direction) as unambiguous vowels is further evidence that the boundary shift observed here is not the result of a contrast in vowel quality. If anchoring occurred at the level of perceived vowel quality, we would predict that when the anchor is "hood" subjects will use the "hud" label more often, or that when the anchor is "hud" subjects will use the "hood" label more often. This prediction is not borne out in these separate analyses nor in the overall analysis. Rather, the data conform to the predictions of the talker contrast model. When the F_0 of the anchor is high, perception of low F_0 tokens is shifted toward "hud", and when the F_0 of the anchor is low, perception of high F_0 tokens is shifted toward "hood".

Fox (1985) conducted a very similar cross-series anchoring experiment and got very different results. The main difference between his experiment and the present one concerned the stimuli. Fox used two continua from "hid" to "head". In one case, the formant range of the continuum was appropriate for a male talker, in the other for a female talker. The tokens with a relatively high formant range were synthesized with high F_0 and the tokens with a relatively low formant range were synthesized with both high and low F_0. On the other hand, the stimuli used in this experiment formed a continuum from "hood" to "hud" and occupied a formant range which was ambiguous between male and female average values. Both the difference between front and back vowels and overall formant ranges contribute to the discrepancy of results. First, in the back vowel continuum used here F_1 and F_2, are positively correlated across the continuum. F_1 and F_2 both increase from the "hood" endpoint to the "hud" endpoint. In Fox's front vowel continua, F_1 and F_2 were negatively correlated. F_1 increased from "hid" to "head" while F_2 decreased from "hid" to "head".[5] If hearers expect generally higher formants when F_0 increases, it is not clear how a continuum in which F_1 and F_2 are negatively correlated would be handled perceptually. There is some evidence that F_1 is more affected by normalization than is F_2 (Ainsworth, 1975), but it is also likely that when information from F_2 contradicts information from F_1, the F_1 information will be less useful than when F_1 and F_2 are correlated. Second, the continuum used here spanned formant ranges which were ambiguous between male and female values. This, coupled with the correlation of F_1 and F_2, meant that this continuum was very

[5] It should also be noted that F_1 and F_2 were also positively correlated in the stimuli used by Fujisaki & Kawashima (1968).

sensitive to a normalization effect; when experimentally manipulated factors influenced perceptual normalization these manipulations were easily observable in subjects' responses to the vowels from the continuum. It is not clear how a talker contrast effect could have produced a shift of identification in Fox's (1985) "hid"–"head" continua.

5. Conclusion

The results of this investigation provide indirect evidence for a talker contrast effect. Research is currently under way to test for the existence of such an effect more directly. If the interpretation given above is correct and a contrast at the level of perceived talker identity is actually taking place, then only one view of vowel normalization remains tenable. In both of the experiments reported here, vowel identification functions were influenced by context. The effect of context was to increase (or cause?) the vowel normalization effect (i.e. tokens with high F_0 had to have higher formant values to be identified as "hud" and tokens with low F_0 had to have lower formant values to be identified as "hood"). The data of Experiment 2 suggest that the influence of context is at the level of talker quality and not vowel quality. Therefore, we conclude that the vowel normalization effect is influenced by talker quality, or, more generally, that perceptual vowel normalization makes reference to percieved talker identity. Of course, it is necessary to point out that this conclusion is based on vowel identification performance in response to only one vowel continuum ("hood"–"hud"). Therefore, the general validity of these results for other vowel contrasts remains to be shown. Assuming that these results are generally valid for other vowels and other languages, they suggest that the algorithmic approach to vowel normalization which is exemplified by Gerstman (1968), Labonov (1971), Nearey (1978), Disner (1980), Sussman (1986), Syrdal & Gopal (1986), Miller (1989) and others, has left out one crucial variable. The information that hearers use to evaluate vowel quality includes not only acoustically available information (such as vowel spectrum and F_0), but also computed information about the person doing the talking.

The comments and criticisms of John Mullennix, Van Summers, David Pisoni, Neal Johnson, Anthony Bladon and an anonymous reviewer are gratefully acknowledged. Ying Yong Qi wrote an earlier version of code to implement the Patterson filters which I adopted here. Research supported by NIH Training Grant No. NS-07134-11.

References

Ainsworth, W. (1975) Intrinsic and extrinsic factors in vowel judgements. In *Auditory analysis and perception of speech* (G. Fant & M. Tatham, editors). London: Academic Press.

Bladon, R. & Lindblom, B. (1981) Modeling the judgement of vowel quality differences, *Journal of the Acoustical Society of America,* **69**, 1414–1422.

Bladon, R., Henton, C. & Pickering, J. (1984) Towards an auditory theory of speaker normalization, *Language and Communication,* **4**, 59–69.

Chistovich, L. A. (1985) Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America,* **77**, 789–805.

Chistovich, L., Sheikin, R. & Lublinskaja, V. (1979) "Centres of Gravity" and spectral peaks as the determinants of vowel quality. In *Frontiers of speech communication research* (B. Lindblom & S. Öhman, editors). London: Academic Press.

Cooper, W. E. (1974) Adaptation of phonetic feature analyzers for place of articulation. *Journal of the Acoustical Society of America,* **56**, 617–627.

Crowder, R. (1981) The role of auditory memory in speech perception and discrimination. In *The cognitive representation of speech* (T. Myers, J. Laver & J. Anderson, editors), North–Holland, New York.

Delattre, P., Liberman, A., Cooper, F. & Gerstman, L. (1952) An experimental study of the acoustic determinants of vowel colour, *Word*, **8**, 195–210.

Disner, S. F. (1980) Evaluation of vowel normalization procedures, *Journal of the Acoustical Society of America*, **67**, 253–261.

Eimas, P. D. (1963) The relationship between identification and discrimination along speech and nonspeech continua, *Language and Speech*, **6**, 206–217.

Eimas, P. D. & Corbit, J. D. (1973) Selective adaptation of linguistic feature detectors, *Perception and Psychophysics*, **13**, 247–252.

Eimas, P. D., Cooper, W. E. & Corbit, J. D. (1973) Some properties of linguistic feature detectors, *Cognitive Psychology*, **4**, 99–109.

Fletcher, H. & Munson, W. A. (1933) Loudness, its definition, measurement and calculation, *Journal of the Acoustical Society of America*, **5**, 82–108.

Fox, R. (1985) Auditory contrast and speaker quality variation in vowel perception, *Journal of the Acoustical Society of America*, **77**, 1552–1559.

Fry, D., Abramson, A., Eimas, P. D. & Liberman, A. M. (1962) The identification and discrimination of synthetic vowels, *Language and Speech*, **5**, 171–189.

Fujisaki, H. & Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels. *IEEE-AU*, **16**, 73–77.

Fujisaki, H. & Kawashima, T. (1969) On the modes and mechanisms of speech perception. *Annual report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, p. 67–73.

Gerstman, L. (1968) Classification of self-normalized vowels. *IEEE-AU*, **16**, 78–80.

Johnson, K. (1989) Higher formant normalization results from auditory integration of F2 and F3. *Perception and Psychophysics*, **46**, 174–180.

Johnson, K. (in press) The role of perceived speaker identity in F_0 normalization of vowels, *Journal of the Acoustical Society of America*.

Klatt, D. (1980) Software for a cascade/parallel formant synthesizer, *Journal of the Acoustical Society of America*, **67**, 971–995.

Labonov, B. M. (1971) Classification of Russian vowels spoken by different speakers, *Journal of the Acoustical Society of America*, **49**, 606–608.

Luce, R. (1959) *Individual choice behavior*, New York: Wiley.

Miller, J. (1989) Auditory-perceptual interpretation of the vowel, *Journal of the Acoustical Society of America*, **85**, 2114–2134.

Miller, R. (1953) Auditory tests with synthetic vowels, *Journal of the Acoustical Society of America*, **25**, 114–121.

Moore, B. C. J. & Glasberg, B. R. (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, *Journal of the Acoustical Society of America*, **74**, 750–753.

Mullennix, J., Pisoni, D. & Martin, C. (1989) Some effects of talker variability on spoken word recognition, *Journal of the Acoustical Society of America*, **85**, 365–378.

Nearey, T. M. (1978) *Phonetic feature systems for vowels*. Bloomington, Indiana: IU Linguistics Club.

Parducci, A. (1965) Category judgment: A range-frequency model, *Psychological Review*, **72**, 407–418.

Parducci, A. (1975) Contextual effects: A range-frequency analysis. In *Handbook of perception* (E. C. Carterette & M. P. Friedman, editors), Vol. II. New York: Academic Press.

Patterson, R. D. (1976) Auditory filter shapes derived with noise stimuli, *Journal of the Acoustical Society of America*, **59**, 640–654.

Peterson, G. & Barney, H. (1952) Control methods used in a study of the identification of vowels, *Journal of the Acoustical Society of America*, **24**, 175–184.

Pickles, J. (1988) *An introducion to the physiology of hearing*, 2nd edition. London: Academic Press.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. PhD thesis, University of Michigan.

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels, *Perception and Psychophysics*, **13**, 253–260.

Pisoni, D. B. (1975) Auditory short-term memory and vowel perception, *Memory and Cognition*, **3**, 7–18.

Plomp, R. (1976) *Aspects of Tone Sensation: A psychophysical study*. London: Academic Press.

Sachs, M. & Young, E. (1979) Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate, *Journal of the Acoustical Society of America*, **66**, 470–479.

Schroeder, M. R., Atal, B. S. & Hall, J. L. (1979) Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In (B. Lindblom & S. Öhman, editors). *Frontiers of speech communication research*, London: Academic Press.

Simon, H. J. & Studdert–Kennedy, M. (1978) Selective anchoring and adaptation of phonetic and nonphonetic continua, *Journal of the Acoustical Society of America*, **64**, 1338–1357.

Slawson, A. (1968) Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency, *Journal of the Acoustical Society of America,* **43**, 87–101.
Summerfield, A. (1971) Information-processing analyses of perceptual adjustments to source and context variables in speech. PhD thesis, Queen's University of Belfast.
Summerfield, A. & Haggard, M. (1975) Vocal tract normalization as demonstrated by reaction times. In (G. Fant & M. Tatham editors). *Auditory analysis and perception of speech* London: Academic Press.
Sussman, H. (1986) A neuronal model of vowel normalization and representation, *Brain and Language,* **28**, 12–23.
Syrdal, A. & Gopal, H. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels, *Journal of the Acoustical Society of America,* **79**, 1086–1100.
Traunmüller, H. (1981) Perceptual dimension of openness in vowels, *Journal of the Acoustical Society of America,* **69**, 1465–1475.
Traunmüller, H. (1982) Perception of timbre: evidence for spectral resolution bandwidth different from critical band? In *The representation of speech in the peripheral auditory system* (R. Carlson & B. Granström, editors). Amsterdam: Elsevier Biomedical.
Young, E. & Sachs, M. (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers, *Journal of the Acoustical Society of America,* **66**, 1381–1403.