The role of perceived speaker identity in $F0$ normalization of vowels

Keith Johnson
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

(Received 19 June 1989; accepted for publication 23 April 1990)

In the experiments reported here, perceived speaker identity was controlled by manipulating the fundamental frequency ($F0$) range of carrier phrases in which speech tokens were embedded. In the first experiment, words from two "hood"–"hud" continua were synthesized with different $F0$. The words were then embedded in synthetic carrier phrases with intonation contours which reduced perceived speaker identity differences for test items with different $F0$. The results indicated that when perceived speaker identity differences were reduced, the effect of $F0$ on vowel identification was also reduced. Experiment 2 indicated that when items presented in carrier phrases are matched for speaker identity and $F0$ with items in isolation, there is no effect for presentation in a carrier phrase. Experiment 3 involved the presentation of vowels from the "hood"–"hud" continuum in two different intonational contexts which were judged to have been produced by different speakers, even though the $F0$ of the test word was identical in the two contexts. There was a shift in identification as a result of the intonational context which was interpreted as evidence for the role of perceived identity in vowel normalization. Overall, the experiments suggest that perceived speaker identity is a better predictor of vowel normalization effects than is intrinsic $F0$. This indicates that the role of $F0$ in vowel normalization is mediated through perceived speaker identity.

PACS numbers: 43.71.An, 43.71.Es, 43.71.Bp, 43.71.Cq

INTRODUCTION

Of all of the hearer's accomplishments in perceiving speech, one of the most important is the ability to adjust to different talkers. In addition to adjusting to differences in dialect and idiosyncratic aspects of articulation, hearers must adjust to physical differences among talkers. This aspect of adjusting to different talkers, which has been called vocal tract normalization or simply normalization, is the focus of this paper. The results of three experiments which examine the role of perceived speaker identity in fundamental frequency ($F0$) normalization of vowels are reported here. These experiments suggest that the perceived identity of the speaker plays a crucial role in vowel normalization. The first section of the paper is a review of the literature. In the second section the results of two pretests are reported. Next, the normalization experiments are described, and, finally, their bearing on vowel normalization theories is discussed.

I. THEORIES OF VOWEL NORMALIZATION

This review of the literature is organized around two binary distinctions among theories.[1] The first distinguishes between the types of information seen as being utilized in vowel normalization, and the second concerns the role that normalization information plays in perceptual vowel normalization.

Following Ainsworth (1975) and Nearey (1989), we can identify two types of information which are used in perceptual vowel normalization. Syllable internal properties such as $F0$ and the frequencies of higher formants play a role in perceptual normalization (Syrdal and Gopal, 1986; Sussman, 1986; Traunmüller, 1981). This *intrinsic* information is potentially useful in perceptual normalization because it is relatively stable for a given talker (as compared with the first two formants). In addition to intrinsic information, syllable external properties such as vowel formant range in a preceding context are also used in perceptual normalization (Ladefoged and Broadbent, 1957; Ainsworth, 1975; Nearey, 1989). Such *extrinsic* information is also useful for normalization because it provides a frame of reference for the perceptual evaluation of $F1$ and $F2$.

Additionally, in order to describe the different theories of normalization that have been proposed, it is necessary to distinguish two roles that normalization information (intrinsic or extrinsic) may play in perceptual normalization. The two views on the role of intrinsic and extrinsic information in normalization will be called *direct* and *indirect* theories of vowel normalization. In a direct theory, it is posited that normalization information is used directly in the construction of perceptual representations of vowels. For instance, in Syrdal and Gopal's (1986) view the proper representation of a vowel is composed of the Bark differences between the $F1$ and $F0$, $F3$ and $F2$, and $F2$ and $F1$. So, using the classifications outlined here, this is a direct theory; Syrdal and Gopal posit that $F0$ (intrinsic information) is used directly in constructing the representation of the vowel. The alternative to direct theories are theories which posit that normalization information is used indirectly in perceptual normalization. For instance, Nearey's (1978) point vowel normalization scheme involves the use of one vowel of known vowel quality to adjust a frame of reference within which other vowels are interpreted. Such an adjustment of a frame of reference is the hallmark of indirect theories. The

 0001-4966/90/080642-13$00.80

extrinsic information, in this case the formant values of a context vowel, is held to affect the frame of reference which is then used in the interpretation of other vowels.

The oldest theories of vowel normalization exemplify the extreme views that Nearey (1989) called *pure intrinsic* and *pure extrinsic*. Potter and Steinberg (1950) proposed that vowel categories are perceptually defined by the relative pattern of stimulation along the basilar membrane rather than the absolute locations of peaks in the spectrum. "Within limits, a certain spatial pattern of stimulation along the basilar membrane may be identified as a given sound regardless of position along the membrane" (p. 812). This theory focuses on the role of intrinsic information and holds that this information directly affects vowel representations. Other *intrinsic direct* theories have been proposed by Traunmüller (1981), Sussman (1986), Syrdal and Gopal (1986), and Miller (1989). In various ways, these authors have implemented the intrinsic direct approach to vowel normalization.

At the other extreme is the *extrinsic indirect* approach to vowel normalization. This type of theory is described by Nearey (1989) as *pure extrinsic*. Joos (1948) was an early proponent of an extrinsic indirect theory of vowel normalization. He suggested that vowel formant values for a given speaker are interpreted relative to the range of formant values for that speaker. Other researchers who have adopted an extrinsic indirect viewpoint are Ladefoged (1967), Gerstman (1968), Labonov (1971), Nordström and Lindblom (1975), and Nearey (1978). There is a good deal of variation in the type of implementation used by these authors, but the basic orientation is the same. For instance, Nordström and Lindblom (1975) used the average $F3$ of a set of back vowels to estimate the length of the speaker's vocal tract and then used this value to scale $F1$ and $F2$. (Note that scaling the formant values of a vowel is functionally equivalent to adjusting a perceptual frame of reference and is quite different from the role of normalization information in vowel representations posited in intrinsic direct theories.) Bladon *et al.* (1984) also proposed an extrinsic indirect theory, although this is not at first obvious. They proposed that normalization is accomplished by shifting auditory spectra of female vowels down by one Bark before comparing them with templates constructed from male vowels. This scheme uses as normalization information whatever information is necessary to determine the gender of the speaker. Bladon *et al.* were not explicit about this, and so it may be best to characterize their approach as an indirect theory using both intrinsic and extrinsic information.

Although less common than the "pure" theories, other combinations of information and process have been proposed. For instance, Broadbent and Ladefoged (1958) proposed a direct mechanism (adaptation level theory, Helson, 1964) to explain the role of context (extrinsic information) in vowel normalization. In the classifactory scheme adopted here, this theory is an extrinsic direct theory because it posits that vowel representations are directly affected by extrinsic information.

Fujisaki and Kawashima (1968) described what may be interpreted as an intrinsic indirect theory. Their experiments involved the manipulation of intrinsic information exclusively, but they introduced the paper with a discussion of the role of pitch and higher formants in perceived speaker identity. In their view, as opposed to the view of Potter and Steinburg (1950) and others who adopt an intrinsic direct approach, syllable internal information serves as a cue to speaker identity and so affects vowel perception indirectly.

In addition to approaches which focus on a single source of information (intrinsic or extrinsic) and a single perceptual role for that information (direct or indirect), there have also been suggestions involving combinations of information and/or processes. For instance, if the geometric mean of $F0$ used in Miller's (1989) approach is taken from surrounding context as well as from the vowel, this theory would then involve both extrinsic and intrinsic information in a direct normalization process. Another way that both intrinsic and extrinsic information might be used in direct normalization involves the postulation of two processes which modify vowel representations—one, the contextual mechanism proposed by Broadbent and Ladefoged (1958) and, the other, the type of intrinsic direct normalization suggested by Potter and Steinburg (1950).

Peterson (1961) and Ryalls and Lieberman (1982) proposed theories in which both intrinsic and extrinsic normalization information are used indirectly. Peterson placed heavy emphasis on formant ratios, as is typical of proponents of intrinsic direct theories, but he also said, "Only to a first approximation do phonetically equivalent vowels have similar formant ratios" (1961, p. 26). He, therefore, suggested that the hearer's previous experience with speech plays a role in the interpretation of vowel formant ratios, citing explicitly both the work of Ladefoged and Broadbent (1957) and Miller (1953). Ryalls and Lieberman focussed on the role of $F0$ in vowel perception and proposed that both intrinsic and extrinsic $F0$ influence vowel recognition indirectly, by providing information about the speaker. As mentioned above, the theory outlined by Bladon *et al.* (1984) may also rely on both intrinsic and extrinsic information in an indirect process.

Finally, Slawson (1968) found that both perceived vowel quality and perceived musical timbre are influenced by changes in intrinsic fundamental frequency and higher resonances. Since the shifts in perceived vowel quality were not as great as the formant shifts found in speech production data, he suggested that perceptual vowel normalization is accomplished by a combination of the intrinsic direct mechanism found to operate both in speech and music perception, and indirect use of intrinsic information (p. 100).

Empirical data suggest that both intrinsic and extrinsic information are used in vowel normalization. Studies by Miller (1953), Slawson (1968), Fujisaki and Kawashima (1968), Summerfield and Haggard (1975), Ainsworth (1975), Traunmüller (1981), and Nearey (1989) demonstrate a perceptual role for intrinsic information, while the work of Ladefoged and Broadbent (1957), van Bergem *et al.* (1988), Remez *et al.* (1987), Ainsworth (1975), Nearey (1978), and Nearey (1989) demonstrate that extrinsic information is used in perceptual vowel normalization. Studies which combine manipulations of intrinsic and extrinsic in-

formation (Ainsworth, 1975, and Nearey, 1989) have found that extrinsic information (specifically, the vowel formant range indicated by precursor vowels) has a greater effect on perceived vowel quality than does intrinsic information such as $F0$ or higher formants.

The fact that there is evidence that both intrinsic and extrinsic information are used in vowel normalization eliminates the "pure" theories. Perceptual vowel normalization necessarily involves both intrinsic and extrinsic information. What is not yet clear is whether this information is used directly in the construction of vowel representations (intrinsic and extrinsic direct) or indirectly in the perceptual interpretation of primary vowel information (intrinsic and extrinsic indirect) or, further, whether the same information can be used in different ways (for instance, Slawson's, 1968, suggestion that intrinsic information is used both directly and indirectly in vowel normalization).

The experiments reported here were designed to assess the role of perceived speaker identity in vowel normalization. If vowel normalization involves indirect use of normalization information (where $F0$ or other stimulus properties indicate the size of the speaker), then it should be possible to predict the results of vowel normalization experiments on the basis of data concerning the perceived speaker characteristics of the stimuli. Perceived speaker identity in the experiments was manipulated by constructing synthetic phrases which had intonational contours spanning different ranges of $F0$. Vowel tokens from a "hood"–"hud" continuum were then embedded in these phrases. Manipulation of the intonational contours of the phrases allowed for a certain degree of control over perceived speaker identity, by making use of the fact that, in natural speech, hearers are confronted (1) with speech from the same talker in which $F0$ is variable and (2) with speech from different talkers in which $F0$ ranges overlap.

Experiment 1 involved the presentation of tokens from a "hood"–"hud" continuum in carrier phrases with intonational contours which sounded like they had been produced by the same talker, even though the $F0$ of the test word was, in one contour, 150 Hz and, in the other, 100 Hz. The experiment, thus, avoided confounding perceived speaker identity with $F0$. Experiment 2 addressed the question of whether simply embedding test words in an intonational context affects vowel identification behavior independently from the $F0$ or perceived speaker identity of the tokens. This experiment also tested whether it is possible to predict the degree of vowel normalization from perceived speaker identity. Experiment 3 involved the presentation of the "hood"–"hud" continuum in carrier phrases that were judged to have been produced by different talkers, while the $F0$ of the test word was identical in the two phrases.

Two pretests were conducted to determine the perceived speaker characteristics of a series of intonational contours, and the results of the pretests were used to select contours for the vowel identification experiments. The next section reports the results of these pretests.

II. PRETESTS

The stimuli used in experiments 1–3 were designed to contrast perceived speaker identity in certain ways. In order to select intonational contours and token $F0$'s that fulfilled the requirements of the experimental designs, two pretest were conducted to evaluate the perceived speaker identity of a set of intonational contours and token $F0$'s. In the first pretest, subjects were asked to judge whether tokens with different intonational contours (in the case of synthetic carrier phrases) or $F0$ levels (in the case of isolated synthetic syllables) had been produced by the same speaker. A fixed-standard AX-discrimination paradigm was used to elicit these judgments. In addition to this "speaker discrimination" task, another pretest was performed in which subjects were asked to label the synthetic "speaker" of the phrases or syllables according to gender and size.

A. Fixed-standard AX-discrimination task

In the fixed-standard AX-discrimination task, a pair of tokens (either two intonational contours or two test words in isolation) were presented to subjects who were asked to classify the items in the pair as having been produced by the same speaker or a different speaker. The only acoustic parameter which was manipulated in the tokens was $F0$. Carrell (1984) found that both $F0$ and vowel formant range were strong cues for speaker identity. The vowel formants which were used in synthesizing the tokens here were ambiguous between those reported by Peterson and Barney (1952) for men and women. Thus the perceived identity of the speaker was easily modified by changes of $F0$.

B. Method

1. Subjects

Ten undergraduate students (seven males, three females) from Indiana University served as subjects in the experiment. Each subject participated in a single 1-h session and received partial course credit for their participation. They were all native speakers of American English and reported no history of speech or hearing disorders.

2. Materials

Sixteen versions of the phrase "This is hood" were synthesized (using the Klatt, 1980, formant synthesizer). Fifteen of these had falling intonational contours which ended at 15 different $F0$ levels (90–160 Hz in 5-Hz steps). The 16th version had a rising intonational contour which ended at 150 Hz. The rising contour can be transcribed in Pierrehumbert's (1980) system of intonational transcription as L*HH%, and the falling contours as H*LL%.[2] The last word in the phrase had steady-state $F0$, and so the intonational patterns were somewhat stylized. Formants of [ɪ] in "this" and "is" were between the values for men and women reported by Peterson and Barney (1952) ($F1 = 456$, $F2 = 1740$, $F3 = 2796$ Hz). The word "this" was 415 ms long and was clearly the nuclear syllable in the phrase. $F0$ contours for all of the intonational contexts are shown in Fig. 1. Fifteen isolated "hood" tokens were also synthesized with steady-state $F0$ levels from 90 to 160 Hz in 5-Hz steps. All synthesizer parameters for the isolated tokens were identical to those used for the word "hood" in the carrier phrase items.

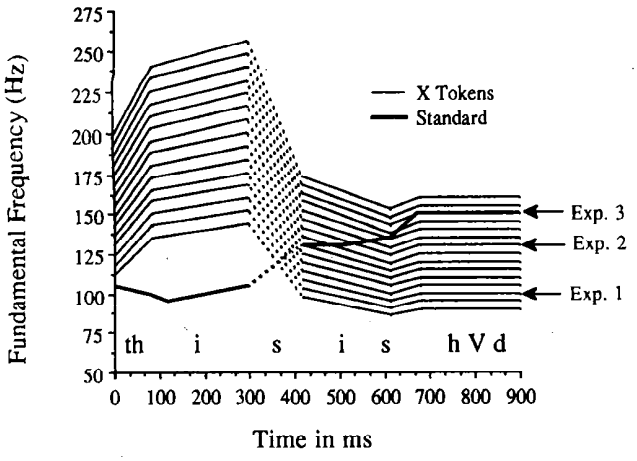

FIG. 1. Intonational contours of "this is hood." The thick line is the rising intonational contour which was used as the fixed-standard in the first pretest. The rising contour was also used in experiments 1–3. The falling contours that were used in experiments 1–3 are labeled.

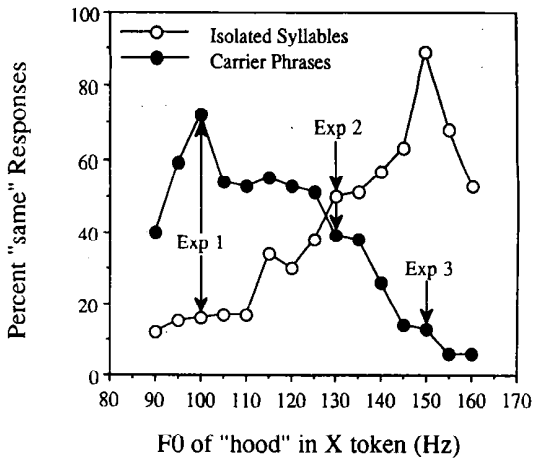

FIG. 2. Results of the fixed-standard AX-discrimination pretest. For both the items in isolation and the intonational contours the fixed-standard token had test word $F0$ of 150 Hz.

3. Procedure

The experiment was conducted on-line on a PDP 11/34 computer at Indiana University. Tokens were played at 10 kHz and low-pass filtered at 4.8 kHz, and presented binaurally over matched and calibrated TDH-39 headphones at a peak SPL of 85 dB. Subjects were run in groups of up to six at a time, and reaction time as well as response data were collected. Labels for the response buttons used by subjects were displayed on a CRT at eye level, and the response to button association was switched for successive blocks of trials.

Subjects were asked to judge the relative identity of the speakers of the tokens in a fixed-standard AX-discrimination task. In this task one token is chosen as a standard against which the other tokens are judged. The subject's task is to decide whether the X token had been produced by the same talker who had produced the A (standard) token. The standard was always presented before the test token. For both the carrier phrases and the isolated syllables, the standard "hood" (in isolation or as the last word in the carrier phrase) had an $F0$ of 150 Hz. The rising intonation contour was the standard for the intonational context items. The items were presented blocked according to context type (carrier phrases versus test words in isolation) and the block of carrier phrases was presented to all subjects before the block of isolated syllables.

C. Results

The speaker discrimination results are shown in Fig. 2. This figure shows the percent "same" responses as a function of the $F0$ of the word "hood" in the X token (the $F0$ of the word "hood" in the A token was always 150 Hz). The isolated syllable tokens are plotted with open symbols, and the carrier phrase tokens are plotted with filled circles. The tokens which were chosen for use in experiments 1–3 are indicated by arrows.

There was a significant interaction between $F0$ level (90–160 Hz) and context (isolated syllables versus carrier phrases) [$F(1,14) = 22.62$, $p < 0.001$]. In addition to this overall pattern, there were some very large differences between subjects which indicate that subjects adopted different strategies for the task. For instance, one subject classified all of the isolated syllables as having been produced by different talkers unless the $F0$ of the item was 150 Hz (identical with the standard). For this subject the task was essentially a pitch matching task at which he was very successful. At the opposite extreme another subject classified all of the isolated syllables as having been produced by the same speaker. For this subject the task may have been one of identifying vocal-tract characteristics or speaker voice source differences which were not present in the tokens. All subjects exhibited the same pattern of results for the carrier phrases (with varying degrees of internal consistency and bias).

The tokens which were chosen for experiments 1–3 had the following properties. The tokens used in experiment 1 contrasted vowels with $F0$ of 100 and 150 Hz in the test word (the item itself for isolated syllables or the last word in the carrier phrase). For isolated syllables, this contrast resulted in a large difference in perceived speaker identity (only 16% "same" responses), while, when the contrast was between a carrier phrase with a rising contour which ended in 150 Hz and a carrier phrase with a falling contour which ended in 100 Hz, the perceived speaker identities were quite similar (72% "same" responses). The tokens used in experiment 2 had roughly equal perceived speaker identity differences. The contrast between isolated syllables with $F0$'s of 150 and 130 Hz (50% "same" responses) was about the same as the contrast between a rising intonational contour which ended at 150 Hz and a falling intonational contour which ended at 130 Hz (39% "same" responses). The carrier phrase tokens used in experiment 3 (isolated tokens were not used in experiment 3) were judged to have been produced by the same speaker only 13% of the time, although they had identical final $F0$ (150 Hz).

III. LABELING EXPERIMENT

To further evaluate the perceived speaker identities of the tokens, a labeling experiment was performed. In this experiment subjects were asked to label each of the tokens used in the fixed-standard AX-discrimination task as either "male" or "female" and as either "big" or "small." Results of this task were then compared with the discrimination task.

A. Method

1. Subjects

Twenty-one undergraduate students (15 males, 6 females) from Indiana University served as subjects in the experiment. Each subject participated in a single 1-h session and received partial course credit for their participation. They were all native speakers of American English and reported no history of speech or hearing disorders. One group of subjects (seven males, four females) labeled the isolated syllables, and one group (eight males, two females) labeled the carrier phrases.

2. Materials

The same synthetic tokens which were used in the previous experiment were used in this experiment.

3. Procedure

The equipment used to run this experiment was the same as that used in the previous experiment. Each item was presented ten times, with order of presentation random. Subjects were asked to label each stimulus as "male" or "female" and then, based on their judgment of the talker's gender, to label the stimulus as "big" or "small." Data were then transformed into a four-point scale (called here a "speaker-size index"): 1 = big male, 2 = small male, 3 = big female, and 4 = small female.

B. Results

Results of the experiment are shown in Fig. 3. In panel (a) speaker-size index scores for the isolated syllables are shown. The horizontal axis is the $F0$ of the "hood" and the vertical axis is the speaker-size index. The solid horizontal line in the graph indicates the score for the item with $F0$ of 150 Hz. The vertical solid lines indicate the degree of difference in perceived speaker identity of the item with 150 Hz and the items with 130 Hz (labeled "Exp 2") and 100 Hz (labeled "Exp 1"). These comparisons are of special interest because tokens with these $F0$ values were used in experiments 1 and 2 described below. In panel (b) speaker-size index scores for the carrier phrases are shown. In this figure, the solid horizontal line is the speaker-size index of the rising contour (the A token in the previous experiment). The comparisons indicated by vertical lines are of special interest because the $F0$ contours being compared were used in experiments 1–3. In general, the labeling data are consistent with the discrimination data presented above. However, there are some discrepancies which merit discussion.

First, note that in the labeling experiment, the carrier phrases were judged to cover a smaller range of possible values of speaker size (ranging from 1.19–2.59) than were the items in isolation (1.05–3.1). This leads to the prediction that in a discrimination task the carrier phrases will have smaller differences in perceived speaker identity than will the items in isolation. If anything, the results of the discrimination experiment show the opposite result, fewer "same" responses when comparing carrier phrases than when isolated syllables. It may be that the full specification of pitch range in the carrier phrases (each item had both high and low $F0$) led to a more fully specified representation of the speaker. This assumption can explain how the carrier phrases could be judged in the labeling task to span a smaller range of speaker sizes while still showing less similarity between the perceived speaker identities in the discrimination task. Note, also, in Fig. 3(b) the tendency for quantal effects. Tokens with final $F0$ from 90–120 Hz form one plateau, tokens with final $F0$ from 125–135 Hz form another, and the rest of the tokens (140–160 Hz) form yet a third quantal region. This slight tendency for the categorical perception of speaker identity is consistent with the assumption that perceived speaker identity is more fully specified by a phrase than by a single item.

Second, data from the labeling experiment [Fig. 3(b)]

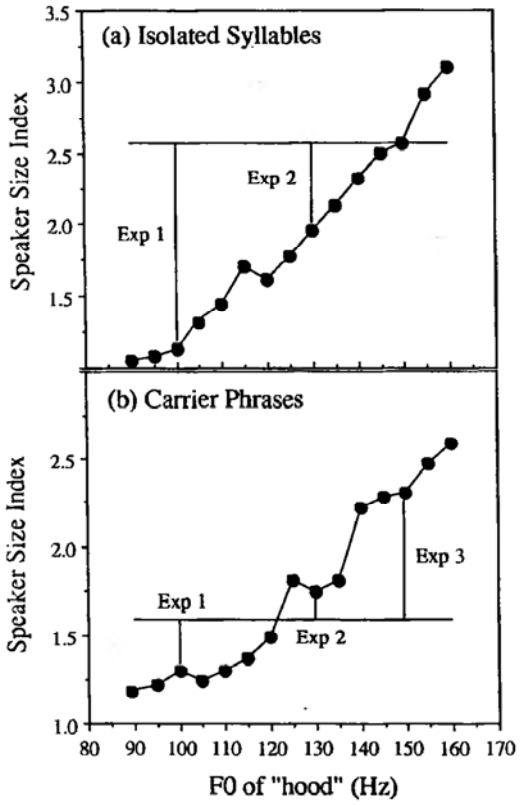

FIG. 3. Results of the labeling pretest. The y axis is an indication of subjects judgments of speaker size. 1 = large male, 2 = small male, 3 = large female, and 4 = small female. (a) Speaker-size judgments for the isolated syllables. (b) Speaker-size judgments for the carrier phrases. In each panel, the solid horizontal line is the score for the token which was used as the fixed-standard in the previous experiment.

do not correctly predict which item with a falling intonational contour would be judged as most similar to the rising intonational contour which was used as the standard in the discrimination task. Based on the labeling data, we would predict that the falling contour which ends with an $F0$ of 120 Hz will be most similar to the rising contour [the horizontal line in Fig. 3(b)], when actually the token judged as most similar to the rising contour was the one that ended at 100 Hz. In general, the discrimination data could be better predicted by the labeling data if the speaker-size index for the rising contour were lower by about 0.25 on the speaker-size index scale. This discrepancy between the labeling and discrimination data may have resulted from a difference in the rate of occurrence of the rising intonational contour in the two experiments. In the discrimination experiment, the rising contour served as the fixed standard and so occurred equally as often as the falling contours (half of the tokens in the experiment had rising contours and half had falling contours). In the labeling experiment, the rising contour occurred once in every 16 trials. This difference in rate of occurrence may have led to a difference in strategy. If, in the labeling task, subjects gave relatively greater weight to the $F0$ of the final word in each contour (a strategy which would have given a correct indication of overall pitch range in 15 out of 16 trials), they would have ended up with higher speaker-size index scores for the rising contour than if they had given equal weight to the entire phrase. Speaker-size scores for the first trial for each token (presumably before the adoption of this strategy) indicate that this explanation is correct. The speaker-size index for the first occurrence of the falling intonational contour which ended at 100 Hz (averaged across ten listeners) was 1.3, while the score for the rising intonational contour (which ended at 150 Hz) was 1.5. Scores from the first trials for the falling contours which ended at 130 and 150 Hz were 2.1 and 2.6, respectively. Thus the discrepancy between the labeling and discrimination data seems to be the result of a difference in listening strategy.

In summary, these pretests have shown that intonational contexts can be used to modify perceived speaker identity. Discrepancies between the voice labeling and discrimination tasks have been attributed to (1) different degrees of speaker specification (contours provide more speaker information than isolated words) and (2) differences in listening strategy which seem to be related to differences in the rate of occurrence of falling and rising intonational contours in the two experiments. Although factors which affect perceived speaker identity are of inherent interest, the experiments discussed here were designed primarily to evaluate the perceived speaker characteristics of the tokens used in a set of vowel normalization experiments. It is to these experiments that we now turn.

IV. EXPERIMENT 1

The fact that there is perceptual evidence indicating that both intrinsic and extrinsic information are involved in perceptual normalization seems to require a model in which both types of information play a role; however, the evidence which has been taken to indicate that intrinsic information is used directly in the construction of vowel representations confounds $F0$ information with speaker information. In all of the studies which have demonstrated an $F0$ normalization effect (e.g., Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968; and Traunmüller, 1981), the vowel sounds to be identified have been presented in isolated syllables with different levels of $F0$. Since the only stimulus variable which changes from one condition to the next is $F0$, it is natural to talk about changes in perceived vowel quality as a function of $F0$. It seems likely, however, that confounded with the differences in $F0$ are also differences in perceived speaker quality because $F0$ is a strong cue for speaker identity (Carrell, 1984, and the pretests above). Thus previous $F0$ normalization experiments do not distinguish between direct and indirect uses of intrinsic information. In this experiment, perceived speaker identity and vowel $F0$ were manipulated separately. If intrinsic $F0$ influences vowel quality directly, we predict that the manipulation of perceived speaker identity will have no effect. On the other hand, if intrinsic $F0$ influences vowel quality indirectly, as a cue to speaker identity, we expect that manipulating perceived speaker identity by means of preceding intonational contour will modify perceived vowel quality.

The $F0$ of synthesized test tokens along a continuum from [hɑd] to [hʌd] was set at 100 Hz for one version of the continuum and 150 Hz for another. The test tokens were presented (in a forced-choice identification task) either in isolation or as the last word in one of two carrier phrases. The phrase "this is hVd" was synthesized with a rising intonational contour (question intonation) or a falling intonational contour (statement intonation). All other synthesis control parameters were identical in the two versions of the carrier phrase. The 150-Hz tokens were presented in the rising intonational context, and the 100-Hz tokens were presented in the falling intonational context. The two carrier phrases had been judged in the first pretest to have been produced by the same speaker 72% of the time, while isolated test words with $F0$ of 100 and 150 Hz were classified in the pretest as having been produced by different speakers. Use of intonational contexts made it possible to separate perceived speaker identity from the $F0$ of the vowel being identified. The items in isolation differed both in $F0$ and in perceived speaker identity and therefore any shift in vowel identification from high $F0$ to low $F0$ could be due to either direct or indirect use of intrinsic $F0$. The items presented in an intonational context differed in $F0$ to the same extent that the items in isolation did, but did not differ as much in terms of the perceived identity of the speaker; therefore, these tokens provide a test of the hypothesis that intrinsic $F0$ is used directly in vowel normalization because $F0$ in these tokens is not confounded with perceived speaker identity.

A. Method

1. Subjects

Twenty undergraduate psychology students at Indiana University (5 males, 15 females) served as subjects in the experiment. They received partial course credit for their par-

ticipation. All subjects reported no history of speech or hearing disorder, were naive to the purpose of the experiment, and participated in a single 1-h session.

2. Materials

A seven-step [hɑd]–[hʌd] continuum was synthesized once with $F0$ of 100 Hz and again with $F0$ at 150 Hz. Formant values of the vowels in the continuum are given in Table I. The endpoints of the vowel continuum had formant values between the average values reported for male [ʌ] and female [ɑ] as reported by Peterson and Barney (1952) and so were ambiguous with regard to speaker sex. The steps in the continuum were equally spaced in Bark.

The [hVd] tokens were synthesized in isolation and in the intonational contexts shown in Fig. 4. The 150-Hz tokens were presented in the rising intonational contour which had been used as the standard in the first pretest, while the 100-Hz tokens were synthesized in the falling intonational context which in the first pretest had been judged to be most similar in perceived speaker identity to the standard. The vowel tokens were steady-state vowels (200 ms) in the consonantal environment [hVd]. Duration of the /h/ portion was 60 ms. The final transitions to /d/ had a duration of 30 ms. Overall duration of the test words was 290 ms. Like the vowel formants of the test tokens, the formant values for [ɪ] in the carrier phrase ("This is ...") were chosen so as to be ambiguous between values typical of males and females ($F1 = 456$, $F2 = 1740$, $F3 = 2796$). "This" was 415 ms long, with a 210-ms steady-state [ɪ]. "Is" was 195 ms long, 90 ms of which was the steady-state vowel. Overall duration of the phrase (including the test word) was 900 ms.

3. Procedure

The experiment was conducted on-line using a PDP 11/34 computer at the Speech Research Laboratory at Indiana University. Tokens were played at 10 kHz and low-pass filtered at 4.8 kHz, and presented binaurally over matched and calibrated TDH-39 headphones at a peak SPL level of 85 dB. Subjects were run in groups of up to six at a time, and reaction time as well as response data were collected. Labels for the response buttons used by subjects were displayed on a CRT at eye level, and the association between response categories and buttons was switched on successive blocks of tokens.

Half of the subjects heard the isolated tokens and half responded to the tokens in intonational context. Each token (or token in its carrier phrase) was presented ten times, and the order of presentation was randomized. A between-subjects design was used with presentation type as the grouping variable. Each subject responded to 140 presentations (7 tokens × 2 $F0$ levels × 10 repetitions).

TABLE I. Formant values of the test tokens used in experiments 1–3.

Token No.	1	2	3	4	5	6	7
$F1$	474	491	509	526	543	561	578
$F2$	1111	1124	1137	1150	1163	1176	1189
$F3$	2416	2424	2432	2440	2448	2456	2464

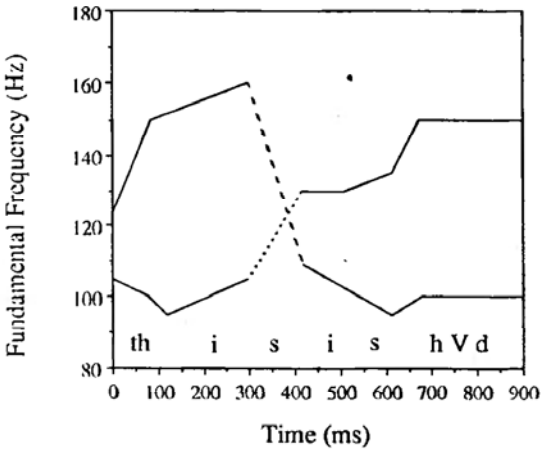

FIG. 4. Intonational contours of the carrier phrases used in experiment 1.

B. Results and discussion[3]

The vowel identification data (Fig. 5) were analyzed in a repeated-measures analysis of variance. The between-subjects factor was PRESENTATION TYPE (items in isolation versus items in intonational context) and the within-subjects factors were TEST WORD $F0$ (high = 150 hz versus low = 100 hz) and TOKEN NUMBER (see Table I). The TOKEN NUMBER and $F0$ main effects were significant [$F(6,108) = 68.53$ and $F(1,18) = 39.06$, respectively, both $p < 0.001$]. The TOKEN × PRESENTATION TYPE and TOKEN × $F0$ interactions were significant [$F(6,108) = 5.15$ and $F(6,108) = 4.87$, respectively, both $p < 0.001$]. As can be seen in Fig. 5, PRESENTATION TYPE and $F0$ level affected the ambiguous middle tokens in the continuum more than they did the continuum endpoints. It is also clear from Fig. 5 that the $F0$ effect was greater when

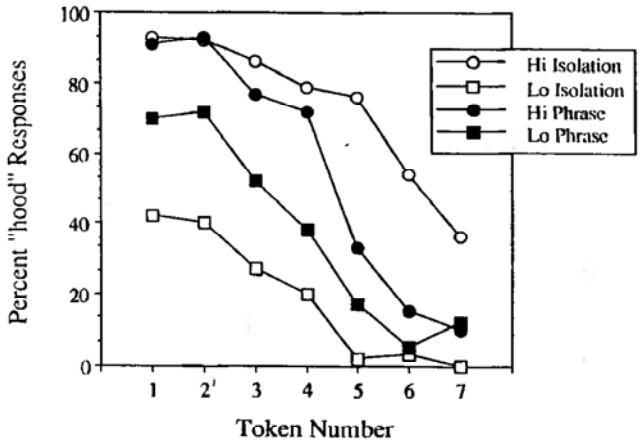

FIG. 5. Results of experiment 1. Identification functions for the high and low $F0$ items in isolated syllables and carrier phrases.

items were presented in isolation than when the test words were presented in context. This interaction (PRESENTATION TYPE×$F0$) was significant [$F(1,18) = 10.04$, $p < 0.01$]. The three way interaction ($F0$×PRESENTATION TYPE×TOKEN) was not significant ($F = 1.37$).

Means (averaged across tokens) and standard deviations of the percent "hood" responses are presented in Table II. The values in this table make it possible to estimate the relative magnitude of an effect which can be attributed to an indirect normalization process. Δ HL is the difference between the high and low $F0$ conditions. As indicated in the table, this difference was reduced when the difference in perceived speaker identity was reduced. Note, however, that in the pretest the two contours used in this experiment were not judged to be the "same" speaker 100% of the time, and so some of the shift seen for the items in carrier phrases may be attributable to some residue of difference in perceived speaker identity. This is also indicated in the results of the speaker identity labeling experiment [see Fig. 3(b)].

The degree of boundary shift for items in carrier phrases was calculated in terms of %$F1$ shift per doubling of $F0$. (There was no 50% crossover boundary in the identification function for isolated syllables with low $F0$, and so no attempt was made to calculate the degree of boundary shift for these items.) Fifty percent crossover boundaries were calculated by linear interpolation from the identification functions for items in carrier phrases in Fig. 5. The boundaries corresponded to $F1$ values of 511 and 536 Hz for the low and high $F0$ continua, respectively. The %$F1$ shift per 100% shift in $F0$ [calculated by formula (1)] was 9.8%. This is on the low end of the 10%–20% per doubling $F0$ reported by Nearey (1989). The fact that the perceptual shift was reduced by reducing perceived speaker identity differences is consistent with the hypothesis that the $F0$ normalization effects reported in previous studies using isolated syllables involved an indirect process similar to the one indicated by these data:

$$[(\Delta F1/F1_{low})/(\Delta F0/F0_{low})]*100. \quad (1)$$

$F1_{low}$ is the lower of two $F1$ boundaries and $F0_{low}$ is the lower of the two $F0$ values used as independent variables in the experiment.

V. EXPERIMENT 2

It is possible that the effect for presentation type found in experiment 1 involves something other than perceived speaker identity. It may be that the perception of a vowel continuum will be constrained by the presence of a preceding context in ways other than the simple perceived speaker identity measure identified in the pretest (Broadbent and Ladefoged, 1958). To investigate this possibility, an experiment was conducted in which the difference in perceived speaker identity scores and the $F0$ values of isolated syllables and carrier phrases were roughly matched. If there is some effect for presentation in a context in addition to these two factors ($F0$ and perceived speaker identity), we would expect this experiment to reveal a difference for these matched phrases and syllables. Experiment 2 also served as a further test of the hypothesis that degree of shift in a normalization experiment can be predicted from perceived speaker identity.

Subjects identified vowels from "hood"–"hud" continua which were synthesized at steady-state $F0$ levels of 130 and 150 Hz. Isolated syllables with these $F0$'s were classified in the first pretest as having been produced by the same speaker 50% of the time. The tokens were also presented in intonational contexts which had been classified (in the pretest) as having been produced by the same speaker 39% of the time. Thus, in this experiment, we expected the contribution of perceived speaker identity difference to be roughly equal for the two presentation conditions, and of course, we expected the contribution of $F0$ differences to be the same whether the tokens are presented in isolation or in a phrase.

Based on the results of experiment 1 the following predictions were possible: (1) Since speaker identities are roughly equal in the two presentation conditions, we predicted that the degree of shift due to changing $F0$ would be roughly the same whether the items are presented in an intonational context or in isolation, and (2) since both presentation types included speaker differences as well as $F0$ differences, we predicted that the degree of shift in $F1$ boundary, as a function of $F0$ doubling, would be greater in this experiment than was the shift in $F1$ boundary for the items in carrier phrases in experiment 1 (where speaker differences were reduced to a minimum).

A. Method

The experimental design and procedure were identical to those in experiment 1. Instead of using tokens synthesized with $F0$ of 100 and 150 Hz (with their respective intonational contours), the tokens used in this experiment were synthesized with $F0$ of 130 and 150 Hz. These intonational contours are shown in Fig. 6. The subjects (8 males and 12 females) were native American English speaking students at Indiana University. They had no reported history of speech or hearing problems. Fifteen of the subjects received partial course credit for their participation, and five were paid a small sum for participating. Ten subjects identified the vowels in isolated syllables, and ten subjects identified the vowels in carrier phrases.

TABLE II. Results of experiment 1. Percent "hood" responses are shown for the high and low $F0$ tokens in isolation and in an intonational context. Standard deviations are shown in parentheses. Δ HL is the difference between the % "hood" responses to the two $F0$ levels, and ΔΔ HL is the difference of the differences.

	$F0 = 150$ Hz	100 Hz	Δ HL
Isolation	73.70 (18.7)	19.14 (17.3)	54.56
Context	55.86 (10.1)	38.00 (10.5)	17.86
		ΔΔ HL	36.70

B. Results and discussion

The results of experiment 2 are shown in Fig. 7. Data were submitted to a repeated-measures ANOVA with between-subjects factor: PRESENTATION TYPE (items

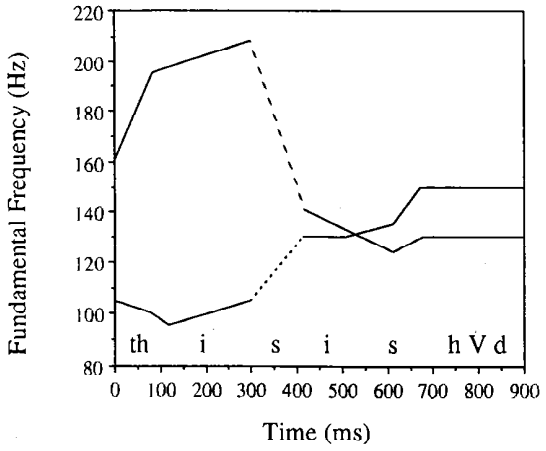

FIG. 6. Intonational contours of the carrier phrases used in experiment 2.

in isolation versus items in carrier phrases) and within-subjects factors: TEST WORD $F0$ (high or low) and TOKEN NUMBER. There was a main effect for TOKEN NUMBER [$F(6,108) = 157.48, p < 0.001$]. The $F0$ main effect and the $F0 \times$TOKEN interaction were also significant [$F(1,18) = 13.82$, $p < 0.01$ and $F(6,108) = 5.42$, $p < 0.001$]. As illustrated in Fig. 7 the $F0 \times$TOKEN interaction occurred because there was a boundary shift in response to the change in $F0$ level. There appears to be a trend for the items in carrier phrases to be identified as "hood" more often than the items in isolation, but this trend was marginally significant [$F(1,18) = 5.48$, $p = 0.031$]. Of greater theoretical interest is the interaction between the $F0$ and PRESENTATION TYPE factors. There was no significant effect of presentation type on the degree of shift in identification when $F0$ was varied [$F(1,18) = 1.22$, $p = 0.284$]. Table III shows these data. Comparing Table III with Table II (results of experiment 1), we see that when differences in perceived speaker identity for the two types of presentation (carrier phrases and isolated syllables) are matched, they produce comparable degrees of shift in the vowel identification functions. Note, however, that the direction of the shift in the carrier phrase condition is actually the opposite of that which would be predicted from the labeling data in Fig. 3(b), although the difference in perceived speaker identity shown in Fig. 3(b) between the rising and falling contours used in this experiment was quite small.

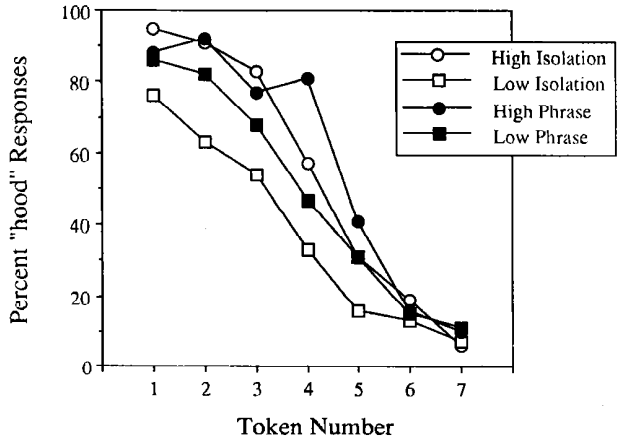

FIG. 7. Results of experiment 2, identification functions for the high and low $F0$ items in isolated syllables and carrier phrases.

TABLE III. Results of experiment 2. Percent "hood" responses are shown for the high and low $F0$ tokens in isolation and in an intonational context. Standard deviations are shown in parentheses. Δ HL is the difference between the % "hood" responses to the two $F0$ levels, and ΔΔ HL is the difference of the differences.

	$F0 = 130$ Hz	100 Hz	Δ HL
Isolation	55.1 (8.9)	37.4 (12.02)	17.7
Context	58.8 (10.1)	48.6 (10.26)	10.23
		ΔΔ HL	7.47

The 50% crossover boundaries in the functions for high and low $F0$ (averaged over presentation type) correspond to values of $F1$ of 535.8 and 517.9 Hz for the high and low $F0$ continua, respectively. Expressed relative to the degree of $F0$ shift in this experiment this $F1$ shift corresponds to a 22.5% increase in $F1$ boundary per doubling of $F0$. This value is markedly greater than the degree of shift found for the items in carrier phrases in experiment 1 and conforms to the predictions of a model of in which intrinsic $F0$ is used indirectly in vowel normalization.

The fact that there was no interaction of presentation type and $F0$ in this experiment indicates that the data in experiment 1 for the items presented in carrier phrases reflect the operation of a process sensitive to perceived speaker identity and not simply an effect of presentation type.

VI. EXPERIMENT 3

In the final experiment, subjects identified vowels from the "hood"–"hud" continuum which were synthesized in carrier phrases which had intonational contours that had identical $F0$ on the test word, even though the items sounded like they had been produced by different speakers (a high-pitched voice making a statement and a low-pitched voice asking a question). The situation created by the use of these carrier phrases is one in which perceived speaker identity varies as a result of the $F0$ of the carrier phrase while test word $F0$ is held constant. If vowel normalization involves only a direct use of intrinsic information (Potter and Steinberg, 1950; Traunmüller, 1981; Sussman, 1986; Syrdal and Gopal, 1986; Miller, 1989), or if the only extrinsic information which can be used in perceptual normalization is the vowel formants of context vowels (Ladefoged and Broadbent, 1958; Ainsworth, 1975; Nearey, 1989), we would expect to find no difference in vowel identification as a result of placing the test words in these contexts. If, on the other hand, extrinsic information is used indirectly in vowel normalization by serving as a cue to the perceived identity of the speaker (in this case by reference to pitch range), we would

predict that there will be a shift in vowel identification such that vowels produced by the higher-pitched voice will be identified more often as "hood" than will the vowels produced by the lower-pitched voice.

A. Method

1. Subjects

Nine (five female, four male) undergraduate psychology students at Indiana University served as subjects. The subjects had not participated in the previous experiments and were naive as to the purposes of the study. They all reported no history of speech or hearing disorder, were native speakers of American English, and received partial course credit for their participation.

2. Materials

The "hood"–"hud" vowel continuum which was used in experiments 1 and 2 was also used in this experiment. The $F0$ of the test word was, in both intonational contexts, 150 Hz. Two different versions of the phrase, "This is hVd," were synthesized. The intonational contours are shown in Fig. 8. In the rising contour $F0$ started at 105 Hz and dipped to 95 Hz during "this" and then rose to 150 Hz by the start of the test word. The falling contour started at 185 Hz and rose to 240 Hz during "this" and then fell to 150 Hz by the start of the test word. In the first pretest, these contours were identified as having been produced by the same speaker only 10% of the time. The falling contour sounded like a talker with a high-pitched voice making a statement, and the rising contour sounded like a talker with a low-pitched voice asking a question.

3. Procedure

The procedure was identical to that used in experiments 1 and 2 except that there was not a manipulation of presentation type. Only items in carrier phrases were presented.

B. Results and discussion

The results of experiment 3 are shown in Fig. 9. A two-way repeated-measures ANOVA was performed on the identification data. The factors were TOKEN NUMBER and INTONATIONAL CONTEXT. As expected, the TOKEN main effect was significant [$F(6,48) = 80.9$, $p < 0.001$]. There was also a significant main effect for INTONATIONAL CONTEXT [$F(1,6) = 13.15$, $p < 0.01$]. Subjects identified items in the falling contour as "hood" 52.5% of the time (s.d. = 5.7), while items in the rising context were labeled "hood" only 44.9% of the time (s.d. = 6.96). Also, the interaction between the TOKEN and CONTEXT factors was significant [$F(6,48) = 5.10$, $p < 0.001$]. As is shown in Fig. 9, the context had an influence on how the ambiguous tokens (Nos. 3 and 4) were labeled, while the perceptual identities of the endpoints remained stable.

The degree of shift in identification in this experiment (7.6% "hood" responses) is smaller than might have been expected from the results of experiment 1 and the first pretest. In experiment 1, the difference between the % "hood" responses to isolated test words and the % "hood" responses to the items in carrier phrases was 36.7%, and since, in the first pretest, the degree of perceived speaker identity difference between the intonational contours used in this experiment was roughly equal to the degree of perceived speaker identity difference between the isolated tokens used in experiment 1, we would expect that the degree of shift in this experiment would be about the same as the reduction of the boundary shift found in experiment 1 ($\Delta\Delta$ HL in Table II). This difference in results could have two explanations. First, it may be that the measure of perceived speaker identity used in the first pretest was not a sufficiently accurate measure of the perceived difference between talkers. Second, it is possible that the tokens in isolation in experiment 1 were subject to a contrast effect which was not present, or not as large, in the carrier phrases.[4]

The data of experiment 3 are directly analogous to the data reported by Ladefoged and Broadbent (1957). How-

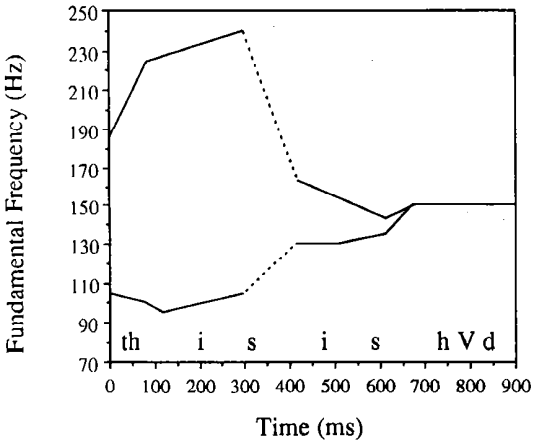

FIG. 8. Intonational contours of the carrier phrases used in experiment 3.

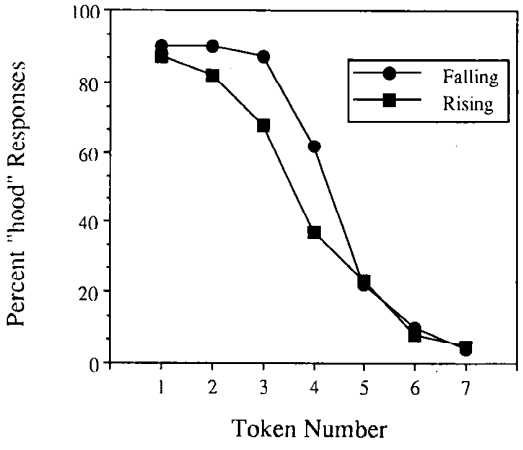

FIG. 9. Vowel identification results for the falling and rising intonation contours in experiment 3.

ever, instead of signaling perceived speaker identity by shifting the formant values of the carrier phrase (as was done by Ladefoged and Broadbent), the identity of the speaker of these phrases was changed by changing the range of $F0$ in the carrier phrase. This difference is important because it indicates that the effect of context in vowel normalization is not direct (involving a vowel contrast mechanism as suggested by Broadbent and Ladefoged, 1958) and points to an indirect involvement of extrinsic information through perceived speaker identity. In the one case, perceived speaker identity information was present in the formant ranges of the carrier phrase, and, in the other case, perceived speaker identity information was present in the $F0$ range of the precursor phrase.

VII. CONCLUSION

A. Summary and implications

In experiment 1, it was found that when differences in perceived speaker identity were reduced (by embedding vowel tokens in carrier phrases), the degree of shift in vowel identification corresponding to a shift in $F0$ was reduced. There was a "residue" shift in the identification functions for the carrier phrase items, which may be due to a direct involvement of $F0$ in vowel representations or to incomplete elimination of differences in perceived speaker identity. The results of this experiment seem to indicate that the role of intrinsic $F0$ in vowel normalization is indirect, serving as a cue for speaker identity. In this interpretation, when the informational content of intrinsic $F0$ is modified by altering the pitch range of a carrier phrase, the impact of that information on vowel normalization is modified.

In experiment 2, it was found that when differences in perceived speaker identity were roughly matched for carrier phrases and isolated syllables, the degree of shift was about the same in the two conditions. This result was taken to indicate that intrinsic and extrinsic $F0$ play the same role in vowel normalization because there was no difference between the perception of vowels in isolated syllables and vowels in carrier phrases when perceived speaker identity was controlled.

Experiment 3 demonstrated that when intrinsic $F0$ is held constant and perceived speaker identity is varied, a shift in vowel identification occurs. The similarity between experiment 3 and the experiment reported by Ladefoged and Broadbent (1957), in view of the fact that vowel formants in the precursor phrases in experiment 3 were not altered, seems to indicate that vowel formant changes in a precursor phrase influence normalization indirectly, through perceived speaker identity.

Overall, the data reported here are consistent with models of vowel normalization in which both intrinsic and extrinsic information play an indirect role. Slawson's (1968) suggestion that intrinsic $F0$ influences vowel categorization both directly and indirectly is also consistent with these data (provided extrinsic information is also taken into account).

In addition to these data, the data reported by Mullennix *et al.* (1989) suggest that intrinsic information is used indirectly in normalization. They found that words were less accurately recognized when the identity of the talker was varied from trial to trial (as opposed to trials blocked by talker). These data can be explained if we assume that hearers construct a representation of the talker in the process of auditory word recognition. This representation would be less accurate in the case where talker identity varies from trial to trial, and thus the probability of incorrect word recognition will increase. (The suggestion, mentioned in the discussion of the pretests above, that carrier phrases more fully specify the identity of the talker is relevant to this argument.)

Also, positing the indirect use of intrinsic $F0$ offers a solution to the "bootstrap" problem (Nearey, 1989, p. 2092). Extrinsic indirect theories involve a circularity. If vowels are perceived relative to a vowel space which is set by context vowels, it is not clear how hearers form a frame of reference for the perception of context vowels in the first place. Extrinsic indirect theories offer no explanation for the perception of context vowels or vowels in isolation. However, if intrinsic $F0$ is used in the construction of a representation of the talker, it can provide a route "into the system" (Nearey, 1989, p. 2092).

B. Generalizability of the results

The experiments presented here have dealt with the vowels [ɑ] and [ʌ] with the assumption that the results apply to vowels in general. Is this assumption valid? It is likely that it would not be possible to replicate these results with a front vowel continuum (for instance, [ɪ] to [ɛ]). In all studies which have found an $F0$ normalization effect using synthetic continua (as opposed to synthetic vowel matrices; see Miller, 1953; Ainsworth, 1975; and Nearey, 1989), the continua have had positively correlated first and second formants. This fact is a further indication that intrinsic $F0$ is used indirectly in vowel normalization. If hearers expect generally higher formants for vowels produced by short vowel tracts, it is not clear how that expectation will affect their perception of a continuum in which $F1$ and $F2$ are negatively correlated across the tokens in the continuum. This does not mean that front vowels are not normalized, just that perceptual normalization is unlikely to produce a consistent effect with a front vowel continuum. This consideration is one constraint on the replicability of the present experiments.

Miller (1953) and Nearey (1989) found that the [ɑ]–[ʌ] boundary was particularly malleable when $F0$ and contextual vowels were manipulated. This malleability may be the result of the fact that these two vowels are very similar in all respects except their formant values. They have similar inherent durations (Peterson and Lehiste, 1960) and formant trajectories (Lehiste and Peterson, 1961). This similarity is less pronounced in other pairs of back vowels (for instance [a] and [ʌ] or [o] and [ʌ]). And the situation is paralleled in the front vowels, where [ɪ] and [ɛ] share many phonetic properties not shared by other front vowels. Therefore, for many vowel pairs, the dynamics of vowel articulation will serve to distinguish between vowels (see Strange, 1989, concerning the confusability of [ɑ] and [ʌ], as well as evidence of individual differences for this confusability). So,

in a sense, the contrast between [ɷ] and [ʌ] is a special case; a special case which makes possible the study of a very interesting property of the human perceptual system.

C. Quantitative predictions of normalization

The experiments here have shown that it is possible to qualitatively predict shifts in vowel identification from differences in perceived speaker identity. In this concluding section, we discuss the quantitative relationship between perceived speaker identity and vowel identification. The data used in this section are drawn from the pretests and the experiments reported here and also from one other experiment which has not been reported. This additional experiment was almost identical to experiment 1 except that the $F0$ values of the test words were 95 and 150 Hz instead of 100 and 150 Hz. Inclusion of data from this experiment increased the number of data points which were entered into the analysis. Correlations between the identification data (magnitude of shift in identification) and the two measures of speaker identity (speaker-size index and speaker similarity judgments) were calculated (Pearson product moment r). These correlations were compared with the correlation between the identification data and the $F0$ differences of the stimuli. These correlations are shown in Table IV.

As indicated in the table, differences along the speaker-size index were most highly correlated with the degree of shift found in the vowel identification experiments (accounting for 82% of the variance). This correlation is shown in Fig. 10. When the differences in $F0$ were used to predict the identification data, only 42% of the variance could be accounted for, and likewise, the correlation between the degree of perceived talker similarity (as measured in the fixed-standard AX-discrimination pretest) and the identification data was low (24% of the variance accounted for). The perceived speaker similarity pretest gave qualitatively correct predictions (as has been emphasized throughout this paper), but it fails to predict the magnitudes of the shifts in identification. The results of experiment 3 are not predicted very well by either measure of perceived speaker identity.

The results of this correlation analysis are consistent with the hypothesis that both intrinsic and extrinsic $F0$ are used indirectly in vowel normalization. It is also clear, however, that quantitative predictions of this sort are very dependent upon the measure of perceived speaker identity employed. Further research is needed to evaluate experimental measures of perceived speaker identity.

In this paper, it has been demonstrated that vowel identification behavior in $F0$ normalization experiments can be both qualitatively and quantitatively predicted from perceived speaker identity. The predictive success of measures of perceived speaker identity suggests that the role of both intrinsic and extrinsic $F0$ in vowel normalization is indirect, as a cue for the identity of the speaker.

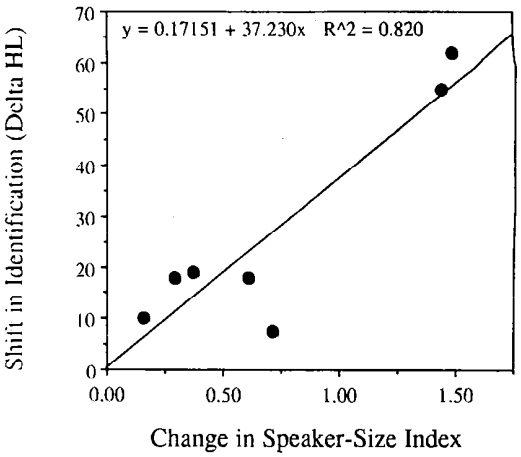

FIG. 10. The correlation between change in speaker size (as measured in the labeling pretest) and magnitude of shift in identification (experiments 1–3 and one other experiment).

TABLE IV. Correlations between shifts in identification and three predictors. Δ HL is magnitude of shift in identification (see Tables II and III). Δ $F0$ is magnitude of change in $F0$ (unreported experiment had $F0$ levels of 95 and 150 Hz). Δ SSI is the difference in perceived speaker identity on the speaker-size index. Percent "different" responses from the first pretest are also shown.

	Δ HL	Δ $F0$	Δ SSI	% "different"
Experiment 1, phrase	17.86	50	0.294	28
isolated	54.56	50	1.433	84
Experiment 2, phrase	10.23	20	0.158	61
isolated	17.70	20	0.608	50
Experiment 3, phrase	7.60	0	0.713	87
95 vs 150 Hz, phrase	19.00	55	0.374	41
isolated	61.70	55	1.480	85
Correlations (r^2)		0.423	0.82	0.24

ACKNOWLEDGMENTS

I appreciate the comments that I have received from Mary Beckman, Rob Fox, Ilse Lehiste, Neal Johnson, John Mullennix, David Pisoni, and Van Summers. I also wish to thank Winifred Strange and Terry Nearey for their careful reviews of the manuscript. The research reported here was supported by NIH Training Grant NS 07134-09 to Indiana University in Bloomington, IN.

[1] This descriptive framework was suggested by T. Nearey in comments on an earlier version of this paper.

[2] In this transcription system H and L stand for relatively high and low $F0$ levels, respectively, and a starred tone indicates a pitch accent (in this case the accent for the nuclear stress on "this"), the percent symbol indicates that the tone is a "boundary tone," and the unmarked H or L is used to indicate a "phrase accent"—a tone which occurs after a nuclear accent and fills up the time until the phrase boundary. The intonation contours used here are the normal patterns for a simple statement (in the case of the falling contour) or question (rising contour) in American English.

[3] Reaction time data were collected, but they do not particularly bear on the hypothesis under consideration. A brief discussion of the general pattern of this data will be given here. This pattern was observed in all three experiments in this study. The reaction time data for the isolated tokens conformed with the pattern found by Pisoni and Tash (1974) (ambiguous items were identified more slowly than unambiguous items). The reaction

times to the items in context were on average faster than to the items in isolation and were undifferentiated across the continuum (no difference between ambiguous and unambiguous items). This conforms to the prediction of Egan (1948) that context serves to allow subjects to prepare for a response. The lack of differentiation between items in the continuum can then be interpreted as a floor effect. Since these considerations do not help in separating the hypotheses under consideration no further analysis of reaction time data was performed.

[4] It is likely that the magnitude of the $F0$ effect for the isolated condition is increased by a contrast effect (Johnson, 1990).

[5] In Johnson (1990) I report the results of two experiments which demonstrate the existence of a quite large contrast effect in $F0$ normalization for tokens presented in isolation and suggest that this contrast effect is a contrast in perceived speaker identity.

Ainsworth, W. (**1975**). "Intrinsic and extrinsic factors in vowel judgments," *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham (Academic, London).

Bladon, A., Henton, C., and Pickering, J. B. (**1984**). "Towards an auditory theory of speaker normalization," Language Commun. **4**, 59–69.

Broadbent, D., and Ladefoged, P. (**1958**). "Vowel judgments and adaptation level," Proc. R. Soc. London (Ser. B) **151**, 384–399.

Carrell, T. (**1984**). "Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification," in Research on Speech Perception Tech. Rep. No. 5, Indiana University, Bloomington, IN.

Egan, J. P. (**1948**). "Articulation testing methods," Laryngoscope **58**, 985–991.

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

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

Helson, H. (**1964**). *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior* (Harper, New York).

Johnson, K. (**1990**). "Contrast and normalization in vowel perception," J. Phon.

Joos, M. (**1948**). "Acoustic phonetics," Language Suppl. **24**, 1–136.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Labonov, B. M. (**1971**). "Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am. **49**, 606–608.

Ladefoged, P. (**1967**). *Three Areas of Experimental Phonetics* (Oxford U.P., London).

Ladefoged, P., and Broadbent, D. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Lehiste, I., and Peterson, G. E. (**1961**). "Transitions, glides and diphthongs," J. Acoust. Soc. Am. **33**, 268–277.

Miller, R. L. (**1953**). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am. **25**, 114–121.

Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.

Mullennix, J. M., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.

Nearey, T. M. (**1978**). *Phonetic Feature Systems for Vowels* (IU Linguistics Club, Bloomington, IN).

Nearey, T. M. (**1989**). "Static, dynamic and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Nordström, P.-E., and Lindblom, B. (**1975**). "A normalization procedure for vowel formant data," in Proceedings of the 8th International Congress of Phonetic Sciences, Leeds, England.

Peterson, G. (**1961**). "Parameters of vowel quality," J. Speech Hear. Res. **4**, 10–29.

Peterson, G., and Barney, H. (**1952**). "Control methods used in a study of the identification of vowels," J. Acoust. Soc. Am. **24**, 175–184.

Peterson, G., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.

Pierrehumbert, J. (**1980**). "The phonology and phonetics of English intonation," unpublished Ph.D. dissertation, MIT.

Pisoni, D. B., and Tash, J. (**1974**). "Reaction times to comparisons within and across phonetic categories," Percept. Psychophys. **15**, 285–290.

Potter, R., and Steinberg, J (**1950**). "Toward the specification of speech," J. Acoust. Soc. Am. **22**, 807–820.

Remez, R., Rubin, P., Nygaard, L., and Howell, W. (**1987**). "Perceptual normalization of vowels produced by sinusoidal voices," J. Exp. Psychol.: Human Percept. Perform. **13**, 40–61.

Ryalls, J., and Lieberman, P. (**1982**). "Fundamental frequency and vowel perception," J. Acoust. Soc. Am. **72**, 1631–1634.

Slawson, A. W. (**1968**). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," J. Acoust. Soc. Am. **43**, 87–101.

Strange, W. (**1989**). "Dynamic specifications of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am. **85**, 2135–2153.

Summerfield, Q., and Haggard, M. (**1975**). "Vocal tract normalization as demonstrated by reaction time," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham (Academic, London), pp. 115–141.

Sussman, H. (**1986**). "A neuronal model of vowel normalization and representation," Brain Lang. **28**, 12–23.

Syrdal, A., and Gopal, H. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.

Traunmüller, H. (**1981**). "Perceptual dimension of openness in vowels," J. Acoust. Soc. Am. **69**, 1465–1475.

van Bergem, D., Pols, L., and Koopmans-van Beinum, F. (**1988**). "Perceptual normalization of the vowels of a man and a child," Speech Commun. **7**, 1–20.