Journal of Phonetics (1997) **25,** 313–342

The perception of intonational emphasis: continuous or categorical?

D. Robert Ladd and Rachel Morton*

Department of Linguistics, University of Edinburgh, George Square, Edinburgh EH8 9LL, U.K.

Received 9th July 1996, and in revised form 3rd April 1997

A series of experiments was carried out to test the idea that there is a categorical difference between ''normal'' and ''emphatic'' accent peaks in English, rather than a continuum of gradually increasing emphasis. This idea builds on serveral studies previously published in this journal as well as a pilot study of our own. The experimental stimuli were all naturally spoken short utterances containing a single rising-falling pitch accent, resynthesised with modified pitch range. In three classical categorical perception experiments we found good evidence of abrupt shifts in identification from normal to emphatic as pitch range increases, but little evidence of an associated peak in discriminability of stimulus pairs. This suggests that the normal/emphatic distinction may be ''categorically interpreted'' but not categorically perceived. Additionally, we report a consistent but puzzling order-of-presentation effect that bears further investigation.

1. Introduction

It is customary to think of an intonation contour as having a linguistically distinctive shape or pattern and an independently variable pitch range. In a one-word English utterance, we may have any one of a handful of distinctive contour shapes—signalling that the contour is, for example, a question or a statement—and any of these shapes may be realized with more or less any pitch range or ''vertical scale''. Some pitch range effects are quite uncontroversially extralinguistic—the differences of overall fundamental frequency (f_0) range due to age and sex differences, for instance—and there can be little doubt that we want to factor these out of the phonetic description of intonation. But even in the case of pitch range effects that convey some kind of linguistic meaning, such as different degrees of emphasis, it still seems appropriate to distinguish them from the shape of the contour, and to treat them as orthogonal (as the ''vertical scale'' metaphor suggests). To take a concrete example, it makes sense to treat all the f_0 contours in Fig. 1 as instances of ''the same'' basic intonation pattern, with variation in the pitch range signalling different degrees of emphasis independently of what is conveyed by the choice of intonation pattern.

* Currently at Entropic Cambridge Research Laboratories, Cambridge, U.K.

0095-4470/97/030313 + 30 $25.00/0/jp970046

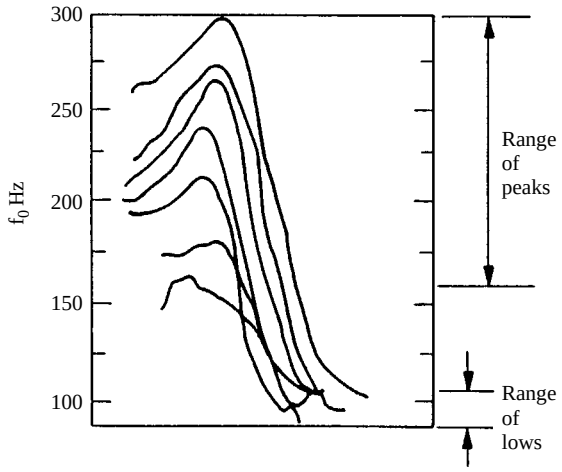

Figure 1. f_0 contour for the same sentence spoken with "the same" intonation in different pitch ranges. Reproduced from Liberman & Pierrehumbert 1984, with permission.

Despite the widespread acceptance of this view, however, and despite its obvious applicability to cases like Fig. 1, it is not without problems. Some of these problems involve cases where it is difficult to decide whether we are dealing with a difference of intonation pattern or a difference of vertical scale. For example, in the British nuclear-tone tradition of intonational description, the difference between a "low-rise" and a "high-rise" nuclear tone is sometimes treated as a difference between two different contour types or patterns (e.g., O'Connor & Arnold, 1973), and sometimes as a difference of vertical scale like any other (e.g., Crystal, 1969). Similarly, in the currently dominant autosegmental/metrical approach to intonational phonology, the issue of distinguishing vertical scale or pitch range effects from phonological distinctions of contour type has played a prominent role in several specific issues of phonological analysis: for example, Pierrehumbert (1980) proposed to treat the difference between downstepping and non-downstepping accents as involving tonally distinct accent types, while Ladd (1983, 1990) argued that downstep is a feature of vertical scale operating on otherwise identical accents.

One general type of case in which there is wide agreement on the appropriateness of a "vertical scale" analysis is that illustrated in Fig. 1, where rising-falling pitch accents have different degrees of emphasis. The shape of the accent contour can be specified independently of the level of the pitch targets (or the interval spanned by the rise and fall), and the degree of emphasis seems to increase gradiently as the vertical scale increases. Such *gradience* is widely supposed to be a characteristic property of pitch range and various other intonational features (cf. Bolinger 1961, Ladd 1994). Even here, however, there have recently been indications that the situation is not straightforward.

The first such indication appears in the study by Gussenhoven & Rietveld (1988). Gussenhoven & Rietveld asked listeners to rate the "prominence" of the second of two accent peaks, in an utterance of the form da-DAH-da-da-da-DAH-da. In a general way they found a close correlation between the pitch range of the second accent (expressed as the height of the second accent peak) and its perceived prominence. This seemed to confirm that pitch range is gradiently variable and

directly signals a gradiently variable meaning like prominence or degree of emphasis. However, Gussenhoven & Rietveld also found that the pitch range of the first accent affects the perception of the prominence of the second in a rather unexpected way. Specifically, they found that if the pitch range on the first accent is reduced, the perceived prominence of the second accent *is reduced as well.* This seems to suggest that prominence is a function of the pitch range of the utterance as a whole, rather than being independently variable to signal the prominence of each individual accent.

Ladd, Verhoeven & Jacobs (1994) replicated and extended Gussenhoven & Rietveld's finding—which they refer to as the Gussenhoven–Rietveld effect—and found that the situation is even more complicated than it had at first appeared. Specifically, they found that the effect occurs only when the second accent peak is not very high. When the second accent peak exceeds a certain level (approximately 145 Hz in their male-voice stimuli), the effect is reversed, and a reduction in the pitch range of the first accent leads to an *increase* in the perceived prominence of the second. This is shown in Fig. 2.

Ladd *et al.* (1994) suggested that the explanation for these findings might "be sought, broadly speaking, in phonology or in psychophysics" (p. 98). They proposed an essentially phonological explanation [developed at greater length in Ladd (1994)], namely that English may have a phonological (and hence categorical) distinction between "normal" and "emphatic" High accents. Given such a distinction, Ladd, Verhoeven & Jacobs explained their experimental results as follows. When accents are normal, the pitch range is evaluated for the phrase as a whole, and we find the Gussenhoven–Rietveld effect. When accents are emphatic, pitch range is evaluated accent-by-accent, and reducing the pitch range of one accent is essentially equivalent to increasing the pitch range of an adjacent one.

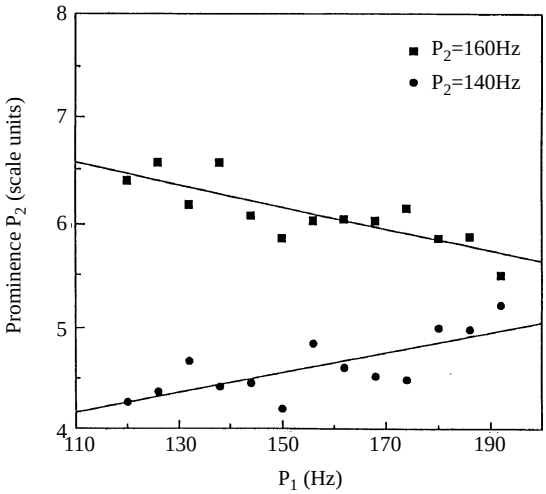

Figure 2. Results from Experiment 1a of Ladd, Verhoeven & Jacobs (1994), showing perceived prominence of a fixed nuclear accent peak P2 as a function of the f_0 of the preceding prenuclear accent peak P1, for two different values of P2. When P2 equals 160 Hz, increases in P1 lead to decreases in the perceived prominence of P2, but when P2 = 140 Hz, increases in P1 lead to *decreases* in the perceived prominence of P2.

Obviously, this explanation depends on the existence of a phonological distinction between normal and emphatic accents. While this runs against the idea of meaningful gradient variability of pitch range, it is by no means out of the question. As Ladd (1994) points out, comparable distinctions between High and Overhigh tone are well attested in African languages. Indeed, the idea of such a distinction in English was suggested by Pike in the 1940s and enjoyed a brief period as part of the accepted American structuralist analysis of intonation (Pike, 1945; Wells, 1945; Trager & Smith, 1951).

Moreover, the idea of such a distinction is consistent with (though not required by) another recent finding, by Hirschberg & Ward (1992), concerning the way listeners interpret the nuance conveyed by a particular intonation contour. Hirschberg & Ward found that the interpretation of the English rise-fall-rise contour (L* + H L− H% in Pierrehumbert's terms) is strongly influenced by acoustic cues to emphasis, especially increased pitch range. Specifically, the rise-fall-rise contour tends to convey what Hirschberg & Ward call ''uncertainty'' when the peak f_0 is low, but ''incredulity'' when the peak f_0 is high. One of Hirschberg & Ward's examples is the following dialogue, with the rise-fall-rise nuclear contour on the word *separating:*

(1) A: I hear John and Mary are calling it quits.
B: They're SEPARATING.

With a normal accent peak on *separating,* this tends to be interpreted as something like ''Well, they're only separating, and they may get back together again''—this is the reading Hirschberg & Ward refer to as ''uncertainty''. With an emphatic accent peak, B's reply tends instead to be interpreted as a surprised question, something like ''Do you really mean to tell me they're separating?''—this is Hirschberg & Ward's ''incredulity'' reading.

The difference between these two interpretations of B's reply seems rather discontinuous, and Hirschberg & Ward's results might be interpreted as evidence for a phonological distinction between normal and emphatic: the emphatic contour (say, L* + H [+emph] L− H%) conveys incredulity, while the normal contour (L* + H [−emph] L− H%) conveys uncertainty. Hirschberg & Ward's example also serves to illustrate a point that is important to keep in mind throughout the following discussion, namely that ''normal'' and ''emphatic'' are simply convenient cover terms for the *phonetic* differences under study. The *pragmatic interpretation* of phonetic ''emphasis'' will vary considerably from one context to another.

If there is a categorical distinction between normal and emphatic accents, we might expect to find evidence of something like a phoneme boundary between the two—for example, an abrupt shift in listeners' perceptions as the pitch range is increased from one side of the boundary to the other. In their study, Hirschberg & Ward used only two rather extreme pitch ranges (i.e., one low and one high), and it is unclear whether we would find any evidence of an abrupt shift as we moved from low to high across a putative boundary level. It is equally possible that the number of incredulity judgements would gradually increase and the number of uncertainty judgements gradually decrease as the range increased from low to high. In that case, pitch range variation could still be seen as an orthogonal dimension, gradient and independent of the phonological representation of the contour, and we would say that the pragmatic interpretation of the single set of phonological specifications L* + H L− H% is influenced by a number of independent variables, such as lexical

choice and gradiently variable pitch range. This is the interpretation of their results—and the consequent prediction—that Hirschberg & Ward themselves would probably favor.

In a preliminary experimental investigation of the nature of the boundary between the normal and emphatic interpretations, Morton (1993) manipulated the pitch range on the phrase *It's Diana again,* spoken with a single accent peak on the word *Diana,* and asked listeners to report their percept of the peak. In some cases, she used a forced choice task (i.e., she asked subjects whether the accent on *Diana* was emphatic or not) and in other cases she used a scalar rating task (i.e., she asked subjects to rate the degree of emphasis of the accent on *Diana* on a ten-point scale). In both cases she found that the response curves appeared S-shaped, i.e., they seemed to reflect a preponderance of non-emphatic judgements at the low end of the stimulus continuum and a fairly abrupt shift to a preponderance of emphatic judgements at the high end. While these results can scarcely be said to establish the existence of a categorical phonological distinction, they are also at odds with the strongest prediction of the ''gradience'' theory.

More systematic investigation based on Morton's pilot experiment seemed warranted. This is the work reported in this paper.

2. General method

Our basic plan for the study involved three distinct stages. Stage 1 was a greatly expanded version of Morton's pilot experiment. The main purpose of this was to see if Morton's results were replicable, and, assuming they were, to select a test utterance for use in Stages 2 and 3. Stage 2 involved two classical ''categorical perception'' experiments of the sort that have been done for segmental phonemic differences (e.g., Liberman, Harris, Hoffman & Griffith, 1957). In this way we aimed to discover whether the normal/emphatic distinction is comparable to other phonological differences such as place of articulation or voicing distinctions in stop consonants. In Stage 3 we planned to extend the findings of Stage 2, looking, for example, for factors that would cause the boundary between perceptual categories to shift. In the event, some modification of Stage 3 became necessary, but the basic structure was preserved.

Several aspects of our method are common to all the experiments and are presented here first. All our experimental stimuli were derived from recorded utterances, here referred to as *source utterances,* which were digitized and resynthesized with modified f_0 contours. This section describes the choice of source utterances and the methods used to create the modified f_0 contours, as well as the procedures for running the experimental sessions.

2.1. *Speech materials*

We began by making recordings of the three sentences *The alarm went off, She's away again,* and *He's Iranian.* The criteria used in choosing these particular sentences were the following. First, they contained a single pitch accent with a

rise-fall intonation pattern. Second, there were no obstruents in the accented syllable which would perturb the f_0 contour. Third, there were no high vowels in the accented syllable, to minimize any effects of intrinsic f_0. Fourth, the three sentences considered together exhibited considerable lexical variety but had a similar basic rhythm of the form *da-da-DAH-da-da.*

Six repetitions of each of the three sentences were recorded by several native speakers of British English. The sentences were presented to the speakers in random order on a set of cards. Three tokens of each sentence were to be read in a ''normal'' voice, and were written in lower case only. In the other three tokens the accented word (*alarm, Iranian, away*) was to be ''emphasized'': the accented word was written in bold capitals and underlined, and the sentence followed by exclamation marks. Recordings were made on digital audio tape (DAT), using professional equipment, in the sound-isolated recording booth in the Phonetics Laboratory of the Edinburgh University Linguistics Department. The recordings were sampled at 16,000 Hz and then analyzed acoustically using the Entropic waves+™ signal processing package.

The utterances of four speakers were chosen for further use. The four were JM (Scottish male), RL (English male—not the first author!), MM (Scottish female), and CS (English female). For each of the four, one ''normal'' and one ''emphatic'' utterance of each sentence were chosen as source utterances (a total of 24 source utterances). All experimental stimuli were created from these source utterances by resynthesizing the utterances with modified f_0 contours, using a pulse-excited linear prediction resynthesis program written by Diego Molla Aliod and Steve Isard. All the tokens chosen as source utterances had relatively smooth pitch contours and produced natural sounding resynthesized speech. In selecting the emphatic source utterances, we also looked for tokens with segmental durations that were as similar as possible to the corresponding normal source utterance. This was not always easy, because some of the speakers tended to lengthen the accented syllable of the emphatic utterances.

2.2. *Modification of pitch range*

As just noted, the stimuli were prepared from the source utterances by modifying the original f_0 values. In this way we created various *stimulus continua* that could be used to test hypotheses about the presence of categorical distinctions in intonation. The most important of these was the peak f_0 continuum. For each speaker, a continuum of 11 peak f_0 values was chosen, ranging from just below the speaker's mean ''normal'' peak value, to approximately the speaker's mean ''emphatic'' peak value. The interval or step size between the peak values on any speaker's continuum was constant in Hz; for RL the step size was 6 Hz, for JM 8 Hz, for MM 10 Hz, and for CS 16 Hz. In addition to the peak continua for each speaker, in Experiment 1 we manipulated various features of the ''prehead'' (the part of the contour on the syllables that precede the accented syllable) independently of the peaks. However, none of these manipulations showed a clear pattern of results, and in the interests of keeping the paper as concise as possible we have decided to omit the rather complex details of these manipulations and any discussion of the results.

For reasons that will become clear later, in different experiments we used two different methods of modifying the pitch range. We refer to these two methods as

the Total Rescaling (TR) method and the Straight Line (SL) method. In the TR method, each individual point in the f_0 contour of the source utterance (i.e., the f_0 value for each analysis frame) is rescaled. This method creates stimuli whose f_0 contours preserve the local perturbations and irregularities of the source utterance. The purpose of preserving the irregularities was to increase the naturalness of the resynthesized speech; work by Silverman (1987) and others suggests that synthetic intonation contours that include local irregularities may be slightly easier for listeners to process than synthetic contours that use smooth line segments.

In the SL method, only certain fixed target points (onset of voicing, peak f_0, etc.) are rescaled, and the remaining f_0 values are generated by simply interpolating between target points using straight line segments. Pitch contours produced in this way obviously do not maintain the irregularities of the original f_0 contour, and are thus more idealized, or ''stylized'' in the sense of 't Hart, Collier & Cohen (1990). Nevertheless, targets were chosen so that the TR and SL stimuli had contour shapes as similar as possible, and sounded as similar as possible. The advantage of using such contours for our purposes was that they make it possible to manipulate the *alignment* of the target points with the text, not just the vertical scaling. This was relevant for Experiment 3. Figure 3 shows examples of pitch contours generated by both procedures, together with the original normal and emphatic contours for comparison.

Irrespective of the choice of the TR or SL method, modifying the pitch range of f_0 contours requires a quantitative model of pitch range. That is, we cannot simply multiply the original contour values in Hz by a constant factor, or increase or decrease them by an additive constant. Even relatively superficial comparison of ''the same'' contour produced in different pitch ranges (cf. Fig. 1) makes it clear that pitch range modification affects different parts of the contour to different extents. The source of these differences appears to be that there is a fairly fixed ''Floor'', or reference value, low in the speaking range of any given speaker, and that f_0 values closer to the Floor are less affected by pitch range modification than values further away from the Floor. This means that the highest values in a contour are the ones most affected by modifications of overall range. This conclusion emerges from various studies, most notably Liberman & Pierrehumbert (1984), and more recently Shriberg, Ladd, Terken & Stolcke (1996); for a more general discussion see Ladd 1996, Ch. 7. At a first approximation, this effect can be modelled by setting the speaker's Floor to zero on a logarithmic frequency scale, and then treating range modification as multiplication by a constant factor. This is the method we used; details are given in the Appendix.

2.3. *Experimental procedures*

In all experimental sessions subjects were seated at computer terminals in a quiet room that could accommodate up to nine subjects at a time. The terminals were spaced out as much as possible, with no two subjects facing each other. Stimuli were presented over loudspeakers. We chose loudspeakers in preference to headphones because they make the unnaturalness of resynthesized speech less noticeable. Also, since in our experiments the variable being manipulated is f_0, the additional high frequency information that might have come through over headphones is essentially irrelevant to the task.

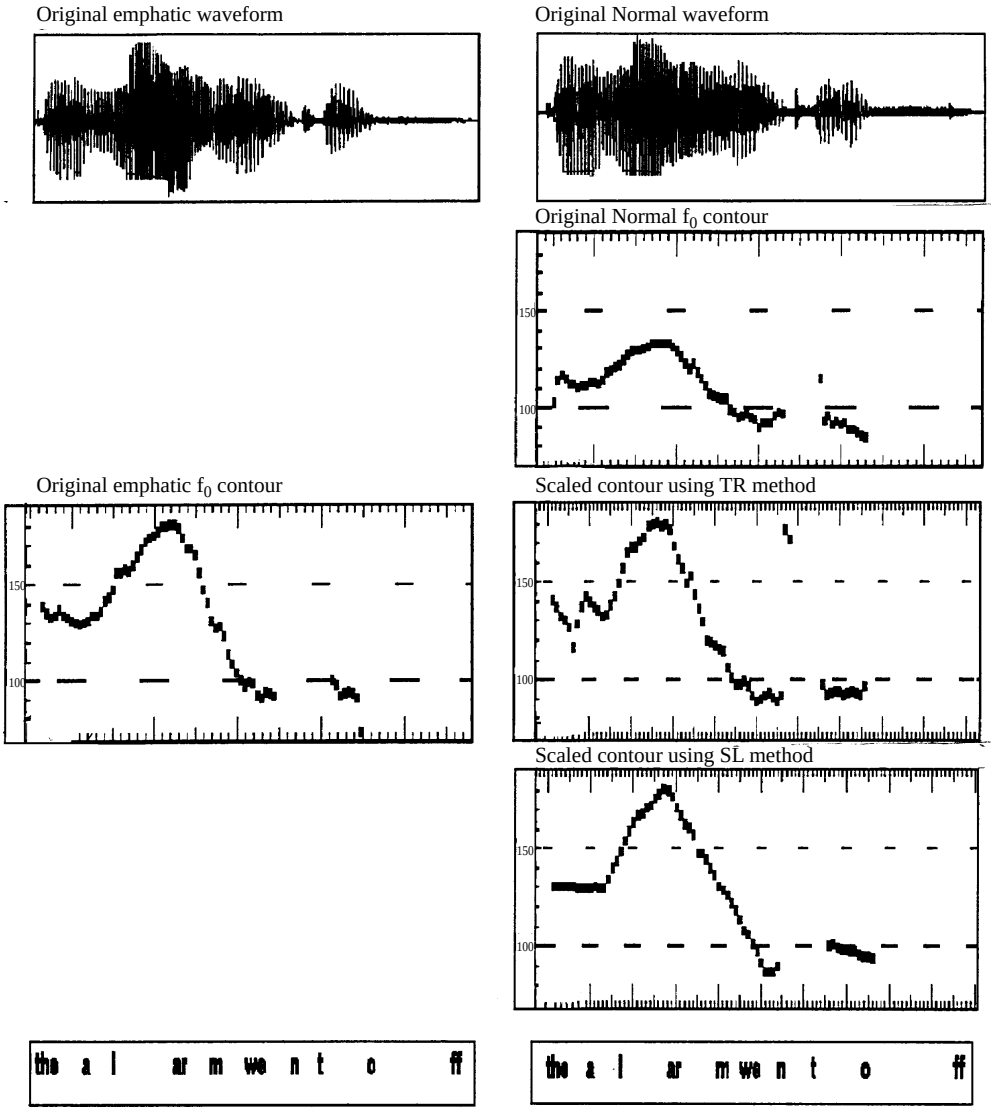

Figure 3. Original "normal" and "emphatic" contours, together with resynthesized "emphatic" contours derived from the normal source contour.

Subjects entered their responses on a keyboard, and saw each response appear on the screen when it was entered. The screen displays were set up according to the task (forced choice or ten-point scale) and in some cases were adapted to specific experiments in other details as well. If subjects pressed an inappropriate key, nothing would appear on the screen; if they wanted to change a response within the interval between stimuli, they simply pressed the key corresponding to the new response. Before each experimental session, typed instructions were given which included examples of the screen display, and subjects were given time to practice using the system.

Each new stimulus was preceded by a warning tone. After the warning tone, the

experimenter cleared all responses for the previous stimulus from all subjects' screens, leaving empty screen displays ready for the next response.

3. Experiment 1

The most important goal of Experiment 1 was to replicate the S-shaped response functions found in Morton's pilot experiment. A preliminary attempt to do this—which we might call ''Experiment 0''—was somewhat discouraging. Where Morton had used only a single speaker and a single sentence (*It's Diana again*), in Experiment 0 we used the three different sentences listed above, and three of the four speakers (MM had not yet been recorded). We also used only the rating scale task (rating degree of emphasis on a ten-point scale) rather than a forced choice task, in order to minimize the possibility that discontinuities or steps in the data were an artifact of the task. Under these experimental conditions, there was no striking confirmation of Morton's results. On the contrary, the overall pattern of data could be taken as broadly consistent with the standard ''gradience'' view, i.e., that the average rated degree of emphasis gradually increases with pitch range.

Nevertheless, for certain combinations of speaker and sentence we did find step-like discontinuities in the ratings, which might reflect the existence of perceptual shifts from ''normal'' to ''emphatic''. On the basis of this rather inconclusive finding, therefore, we ran a more general exploratory experiment, whose purpose was the systematic comparison of different combinations of conditions on the linearity or otherwise of the responses. We hoped to identify certain combinations that would favor clearly S-shaped response curves, in order to use one such combination as the basis for a classical categorical perception study, in accordance with the overall plan of the investigation outlined in Section 2. Because of the exploratory nature of this experiment, we manipulated several variables, both acoustic and otherwise.

3.1. *Method*

3.1.1. *Experimental variables*

As in Morton's pilot experiment, the central stimulus variable was peak f_0, which had 11 levels for each combination of other stimulus variables. We used two sentences (*He's Iranian* and *The alarm went off*) and the four speakers listed above. The exploratory variables included Source (normal *vs.* emphatic), Prehead (smooth *vs.* irregular), Speaker accent (Scottish *vs.* English), Subject accent (Scottish *vs.* English), and Instruction (''linguistic'' *vs*. ''paralinguistic''). Only Source and Instruction produced clear results and these are the only two we report in any detail.

Source was manipulated by choosing two source utterances for each speaker-sentence combination: one originally spoken as ''normal'' and the other originally spoken as ''emphatic'' (cf. Section 2.1.). For each normal/emphatic pair of source utterances we adjusted the source f_0 contours before rescaling, so that the stimuli were as close to identical as possible, while still preserving the local perturbations and irregularities of the original utterances. However, we made no adjustment for the fact that the Peak of the emphatic contours was invariably aligned later in the

accented syllable than the peak of the normal contours. For example, in RL's *alarm* sentences, the f_0 peak is reached about two-thirds of the way through the vowel [a:] in the normal source utterance, but not until late in the nasal [m] in the emphatic source utterance. This later alignment appears to be a rather general property of emphatic high accents in English and some other languages (cf. Ladd, 1983; Kohler, 1987). For purposes of Experiment 1b, this alignment difference was in effect treated as part of a global package of differences between the two types of source utterance, primarily because we could not control the alignment of the peak using the TR method of creating the experimental contours. We were, of course, aware that peak alignment might be a significant variable in itself, which was why we tried the SL method of stimulus creation in Experiments 2b and 3.

The other variable of interest here was Instruction. Two different types of instructions to subjects were used. In the first, called here "paralinguistic" instructions, subjects were asked to rate the degree of emphasis on each utterance on a ten-point scale, which was the approach we had used in the inconclusive Experiment 0. We thought that this type of instruction might bias subjects to interpret the stimuli in a more linear or continuous way, by focusing their attention on the global, paralinguistic aspects of raising the voice, and that if we instruct subjects to focus on a single accented word, we bias them to interpret the stimuli in a more linguistic or categorical way. We therefore included a second type of instruction, here called the "linguistic" instruction, which was intended to resemble the instruction of Morton's pilot experiment more closely. For each sentence, we devised two contexts in which the sentence might be uttered, one appropriate for a normal or unemphatic reading, and the other for an emphatic or contrastive reading. For the sentence *He's Iranian,* the utterance could be "just a neutral statement about where he's from" or "a contradiction or correction" (e.g., He's not Armenian). For the sentence *The alarm went off,* the suggested paraphrases in the "linguistic" instruction were "it was an everyday occurrence" and "it was an unusual or frightening experience". Subjects were asked to judge, again on a ten-point scale, which interpretation was the more likely intended meaning of the stimulus they had just heard. This procedure is reminiscent of that used by Hirschberg & Ward (1992), in the sense that subjects rated context-specific paraphrases, rather than general dimensions like degree of emphasis. However, we preserved our basic procedures of using a continuum of peak values (rather than two extreme stimuli as in Hirschberg & Ward's study) and a ten-point rating scale (rather than a forced choice), to avoid the possibility that any "steps" in the judgement curves are simply artifacts of the subjects' task.

3.1.2. *Experimental groups*

Because different contexts were suggested in the instructions for each sentence, the sessions had to be blocked by both Sentence and Instruction type. A different experimental session was run for each of the four possible combinations of Sentence and Instruction (Iranian-Paralinguistic, Iranian-Linguistic, Alarm-Paralinguistic, Alarm-Linguistic), with different subjects in each. There were 35 subjects in total, either eight or nine in each of the four groups. Subjects were mainly first year undergraduates in linguistics fulfilling a course requirement.

Two tapes were made, one for each sentence. Both tapes had a practice session of

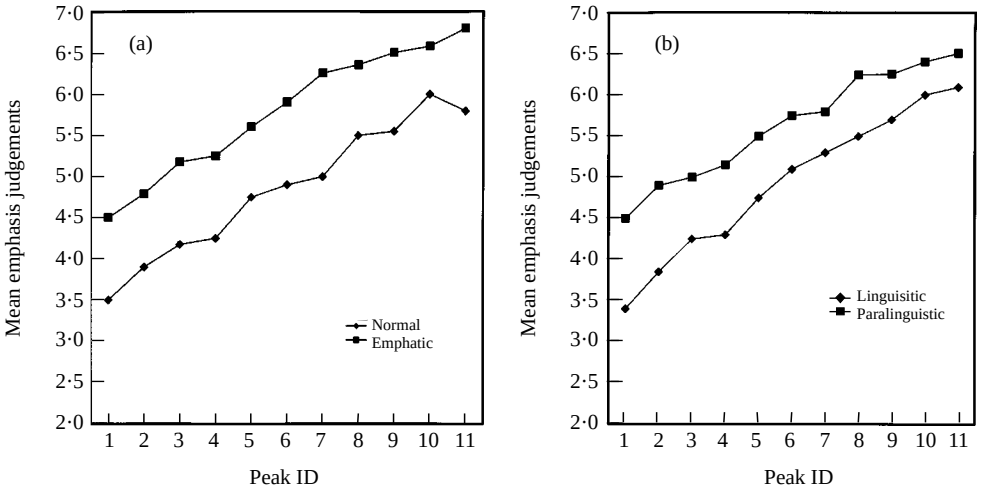

Figure 4. Selected results from Experiment 1. (a) shows the main effect of Source utterance (normal *vs.* emphatic) on the subjects' judgements, and (b) shows the effect of Instruction (linguistic *vs.* paralinguistic).

16 stimuli (four per speaker), followed by one token of each of the 176 (2 source $\times$ 11 Peak $\times$ 2 Prehead Shape $\times$ 4 Speakers) stimuli. The stimuli were randomized, but the random order was corrected to remove sequences of three or more stimuli with the same Peak f_0, Prehead Shape and/or Source. The same order of stimuli was used for both linguistic and paralinguistic sessions to avoid confounding any effect of Instruction.

3.2. *Results*

Data were analyzed in a series of ANOVAs. Because there are so many variables and because this part of the study was exploratory, many details of these analyses are omitted. However, two clear results are worth highlighting. First, as can be inferred from Fig. 4(a), there is a very strong main effect for Source [$F(1, 27) = 73.07$, $p < 0.0001$]. Stimuli derived from emphatic source utterances are rated as more emphatic than those derived from normal source utterances. Source is also involved in a great many interactions. This strongly suggests that a variety of acoustic cues to emphasis survive the f_0 manipulation. (More mundanely but equally importantly, it also suggests that the rating scale task is valid.) Second, there is a main effect for Instruction [$F(1, 27) = 7.75$, $p < 0.01$], with the paralinguistic instruction yielding more emphatic judgements than the linguistic instruction, and a Peak $\times$ Instruction interaction [$F(1, 270) = 2.88$, $p < 0.05$], with Instruction having a larger effect at lower (i.e., less "emphatic") Peak values [see Fig. 4(b)]. Apart from the main effect and this one interaction, Instruction shows up only in one 4-way and two 5-way interactions that are essentially uninterpretable.

For purposes of the present paper, our primary goal in Experiment 1 was to identify some combination or combinations of variables that yielded clear steps in the response curves, in order that we might use these in preparing stimuli for the categorical perception studies. Several combinations of variables led to response

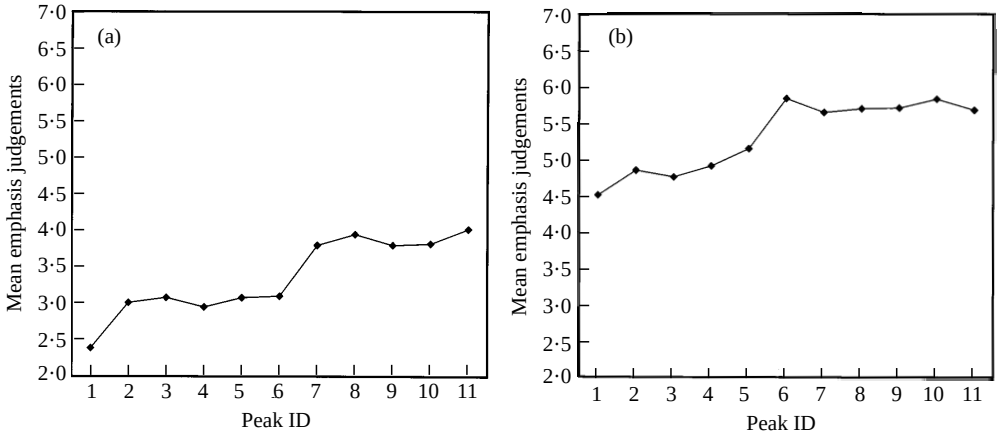

Figure 5. Stepping response curves from Experiment 1. (a) Speaker RL, normal Source, *alarm* sentence; (b) Speaker MM, normal Source, *alarm* sentence.

functions that included such abrupt steps. Results from two of these combiantions are plotted in Fig. 5.

For use in subsequent experiments, we selected the following combination of variables: Speaker RL, normal source, linguistic instruction, smooth prehead, and *alarm* sentence [Fig. 5(a)]. We chose this combination for three reasons. First, more is known, both formally and informally, about male pitch ranges and male voices than about female ones, and we felt that by using a male voice we would be more likely to avoid unanticipated problems in creating stimuli. Second, RL had the most standard, least regionally marked accent of all four speakers, and again, we felt that we might avoid pitfalls by using a standard accent. Third, RL's results seemed most internally consistent overall, in particular yielding more clear main effects and fewer hard-to-interpret interactions than the other speakers. We are aware that all three of these reasons involve little more than hunches, and that the first two serve to perpetuate the supposed neutrality of male voices with standard accents; we report them here for the sake of completeness.

4. Experiment 2: categorical perception

The best documented and established methodology for investigating the existence of categories is that of categorical perception (CP). In the classical CP paradigm used to investigate category boundaries between segmental phonemes, two tasks are involved. The first is a forced choice identification task, in which subjects are presented with stimuli from an acoustic continuum and asked to identify them as a member of one of two categories. In the second task, subjects are asked to discriminate between pairs of stimuli taken from the continuum used in the identification task. If perception is categorical, then discrimination between stimuli at the category boundary is predicted to be better than between stimuli identified as members of the same category. In the original work by Liberman *et al.* (1957), a formula is given for predicting discrimination results from identification results. We dispense with the mathematical details here, but emphasize that, for classical CP, it

is not enough to observe an abrupt shift in the identification function: such a shift must be accompanied by a peak in the discrimination function. (For more discussion of CP see the papers in Harnad (1987).)

We conducted two CP experiments based on the material investigated in Experiment 1. In both experiments we used one of the combinations of stimulus and subject variables that had emerged from Experiment 1 as most likely to involve a discontinuity between normal and emphatic: speaker RL, normal source utterance, smooth prehead, *alarm* sentence, and linguistic instruction. The aim was to investigate whether the discontinuities that appeared in the results of Experiment 1 involve the sort of CP frequently found with phoneme categories.

4.1. *Experiment 2a*

4.1.1. *Method*

4.1.1.1. *Identification task.* Because we had selected a single speaker, sentence, prehead, and instruction for use in these experiments, there were only 11 different stimulus types, one for each step on the continuum of peak f_0. The materials for the identification task consisted of 10 repetititons of each of the 11 stimuli. The stimuli were recorded onto a tape in random order, with a 3 s interstimulus interval that included a 1 s warning tone. Subjects gave a forced choice response after each stimulus by pressing one of two buttons. The labels ''everyday occurrence'' and ''unusual experience'', which had been used in Experiment 1 to label the poles of the rating scale in the linguistic instruction for the *alarm* sentence, were shown on the screen display above the response boxes.

4.1.1.2. *Discrimination task.* The materials for the discrimination task consisted of pairs of stimuli from the identification task. Stimuli were presented in a so-called 2IAX format, in which subjects are presented with two stimuli in quick succession, and then asked to state whether the two are the same or different. This task is simple to explain to subjects, and invovles less short-term memory burden than other discrimination tasks. (For a full discussion of various discrimination tasks see Repp, 1984).

The stimuli in the discrimination pairs were one step apart on Speaker RL's Peak f_0 continuum, i.e., their peak levels differed by 6 Hz; it appeared on the basis of a pilot experiment that two-step pairs (with peak levels 12 Hz apart) might be too easily discriminated. There were thus a total of ten stimulus pairs: step 1 + 2, step 2 + 3, and so on up to step 10 + 11. These stimulus pairs, in which the two stimuli were different from each other, will be referred to as AB pairs. In addition to the 10 AB pairs, 10 control AA pairs were created, which contained two identical stimuli (step 1 + 1, etc., excluding step 11 + 11). Ten repetitions of all 20 stimulus pairs were randomized and recorded onto tape. The interval between the two members of each stimulus pair was 1 second. The interval between pairs, as in the other experiments, was 3 seconds, which included a warning tone.

Subjects were asked to decide whether the stimuli in each pair were the same or different, by pressing one of two labelled buttons on the keyboard. The screen display showed the labels ''same'' and ''different'' above the appropriate box.

4.1.1.3. *Experimental procedures.* There were two experimental sessions with a total of 12 subjects participating. As is customary, the identification task preceded the

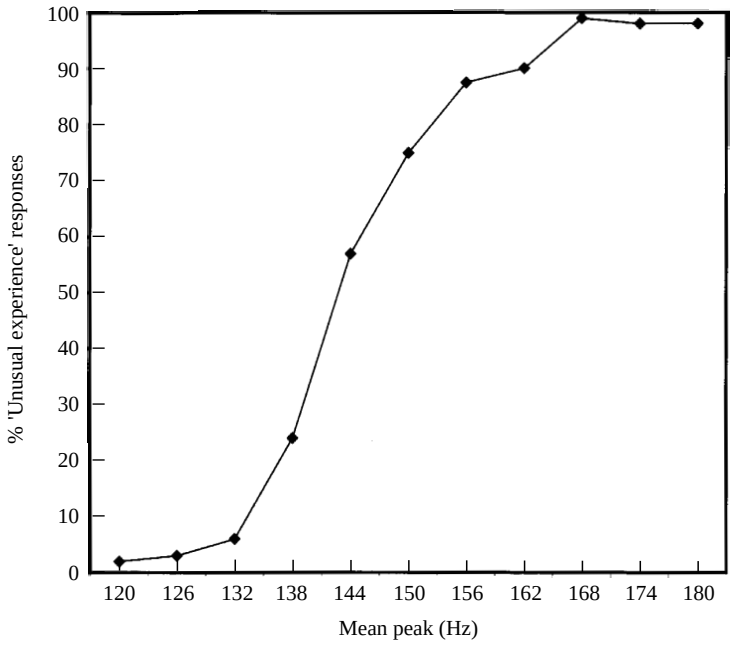

Figure 6. Results of identification task, Experiment 2a.

discrimination task. The main experimental session was preceded by a practice session, in which subjects heard five tokens of step 1 (120 Hz peak) and five tokens of step 11 (180 Hz peak), in random order, so that they might become acquainted with the speaker's pitch range, and have some practice identifying the categories.

The 110 stimuli in the identification task were presented in two blocks of 55 each, with a short break in between blocks. At the end of the identification task there was a slightly longer break before the discrimination task. The 200 stimuli of the discrimination task were presented in four blocks of 50 stimuli, with a short break in between each block. The whole experiment lasted roughly 1 h.

4.1.2. *Results*

4.1.2.1. *Identification task.* Results of the identification task are plotted in Fig. 6. The response curve appears strongly S-shaped. The range from 120 Hz to 132 Hz appears to correspond to an "everyday occurrence" category, while the range from 168 Hz to 180 Hz corresponds to an "unusual experience" category: within these two ranges the percentage of "unusual experience" responses is close to 0% and 100% respectively, and there is a steep shift from one response to the other in the range values between 138 Hz and 162 Hz.

Probit analysis (Finney, 1971), used to calculate the boundary between categories in sigmoid response curves, suggested that the threshold of the "unusual experience" category was at 144.9 Hz. In order to show that the response curve is genuinely S-shaped rather than linear, regression lines were fitted to both the identification data (i.e., the percentages of "unusual" responses for each step on the continuum) and to their transformed probit values. If the identification function is in fact S-shaped, the probit transformation will straighten it out, and a straight

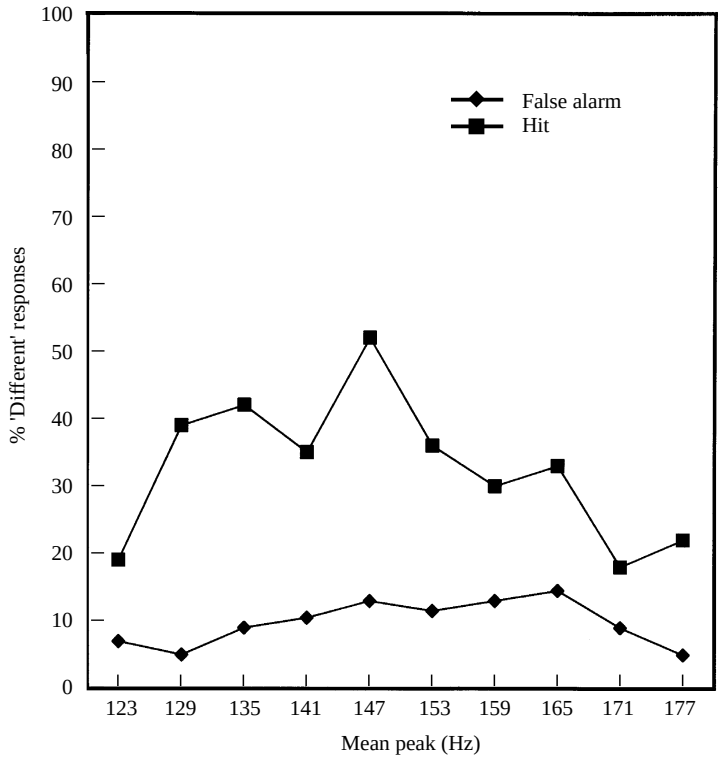

Figure 7. Results of discrimination task, Experiment 2a.

regression line will therefore fit the transformed data more closely than the original data. This is exactly what happens ($R^2 = 0.89934$ for the untransformed data; $R^2 = 0.94489$ for the transformed data), though the difference is admittedly small.

4.1.2.2. *Discrimination task results.* Figure 7 shows the total percentage of "hits" and the total percentage of "false alarms" plotted against the mean peak f_0 of the stimuli in each pair along the Peak continuum. "Hits" are "Different" responses to stimulus pairs that are actually different, i.e., AB pairs. "False alarms" are "Different" responses to stimulus pairs that are in fact the same, i.e., AA pairs.

There does appear to be a discrimination peak (i.e., a peak in the "hit" rate) at 147 Hz, which is approximately what is expected from the identification function. Even when the false alarm rate is subtracted from the hit rate, the peak remains in the same place. This result lends some credence to the idea that there is a categorically perceived boundary between normal and emphatic peaks. Moreover, the f_0 of the putative boundary—between 145 and 150 Hz—is precisely consistent with the findings of Ladd *et al.* (1994). However, there are serious reasons for treating the result with caution.

First, the overall discrimination rate is very low. While the plot of "hits" is distinct from the "false alarm" plot, showing that subjects responded differently to AB and AA pairs, the highest percentage of hits does not even reach 50%. This suggests that subjects had difficulty hearing the 6 Hz differences in the AB pairs. Yet as noted above, a pilot experiment suggested that a 12 Hz difference would have been too easily perceptible. This squares poorly with the idea that there is a clear difference

between within-category and across-category perception, which is the essence of CP. Second, the false alarm rate neither remains flat nor shows a gentle peak at the putative category boundary, which is what we would expect on the basis of classical CP phoneme boundary studies. Instead, the rate steadily increases with peak f_0 up to step 8 + 9, after which it drops slightly. This overall trend suggests a response bias that gradually increases with pitch range, i.e., subjects are more likely to think that stimulus pairs are different the higher the overall pitch. If this is true, it seems more consistent with the idea of gradiently variable pitch range rather than with a physical continuum divided up into phonological categories. Note, however, that this trend in the response bias is the opposite of what would be predicted on the basis of a logarithmic pitch scale (as used by e.g., Thorsen, 1980; 't Hart *et al.,* 1990). While the appropriate psychophysical scale for speech pitch is still a matter of considerable uncertainty, many investigators agree that pitch perception is not linear. (For example, the ERB scale, proposed by Hermes & van Gestel (1991) on the basis of psychophysical work by Glasberg & Moore (1990), is roughly midway between linear and logarithmic.) Obviously, if this is the case, stimulus pairs that are equal intervals apart on a linear Hz scale should be *harder* to distinguish as the f_0 increases. We return to this phenomenon in more detail in Section 6.3.

4.1.3. *Discussion*

The results of Experiment 2a are somewhat encouraging from the perspective of our original idea; the identification task gives results that are clearly in line with the hypothesis of a category boundary between normal and emphatic, and there is a modest discrimination peak at the boundary determined by probit analysis of the identification results. At the same time, however, the discrimination within the apparent categories is still higher than zero, suggesting that subjects could hear pitch differences within categories, while discrimination at the boundary scarcely reaches 50%, which at the very least shows that subjects could not do the task very well. Moreover, the steady increase in the false alarm rate is difficult to reconcile with standard findings of CP experiments. A replication experiment was therefore essential.

4.2. *Experiment 2b*

Experiment 2b had three distinct purposes. The first was to replicate Experiment 2a, specifically looking to see if we could again find evidence for the conjunction of a discrimination peak with an identification boundary. The second purpose was to check the possibility of using the Straight Line (SL) method of contour generation: this was a necessary prerequisite to Experiment 3, as will be seen below. The third reason for the experiment was essentially a matter of methodological rigor, but turned out to have the most interesting repercussions.

In Experiment 2a, the stimulus pairs in the discrimination task were, as noted above, either AB pairs (step 1 + 2, step 5 + 6, etc.) or AA pairs (step 1 + 1, etc.). Primarily because of an oversight—though it was an oversight motivated by the need to keep the experimental sessions to a manageable length—there were no BA pairs, i.e., pairs of the sort step 2 + 1, in which the two stimuli were different but the pitch excursion decreased across the pair rather than increasing. The third motivation for

doing Experiment 2b was to remedy this oversight, and include both AB and BA pairs in the discrimination task, in addition to the identical control pairs.

4.2.1. *Method*

4.2.1.1. *Stimulus preparation.* As just stated, the stimuli in this experiment were prepared using the SL method of pitch range modification rather than the TR method. In all other respects the stimuli were identical to those of Experiment 2a.

4.2.1.2. *Identification task procedure.* This was identical to the identification task in Experiment 2a.

4.2.1.3. *Discrimination task.* This was identical to the discrimination task in Experiment 2a, except (as just noted above) in addition to AB stimulus pairs, in which the second peak of the pair is higher than the first, we also used BA stimulus pairs, in which the second peak of the pair is lower than the first. Each of these pairs was also matched by an AA or BB control pair, the AA series running from step $1+1$ to step $10+10$, and the BB series running from step $2+2$ to step $11+11$. Because the number of stimulus pairs was thus doubled, the number of presentations of each pair had to be halved from 10 to 5 in order to keep the duration of the experimental session under an hour.

4.2.1.4. *Experimental procedures.* The experimental procedures were the same as those used in Experiment 2a. There were two experimental sessions with a total of 13 subjects, most of whom were once again linguistics undergraduates fulfilling a course requirement.

4.2.2. *Results*

4.2.2.1. *Identification task.* The results of the identification task are shown in Fig. 8. Once again we find a strongly S-shaped response curve with a preponderance of

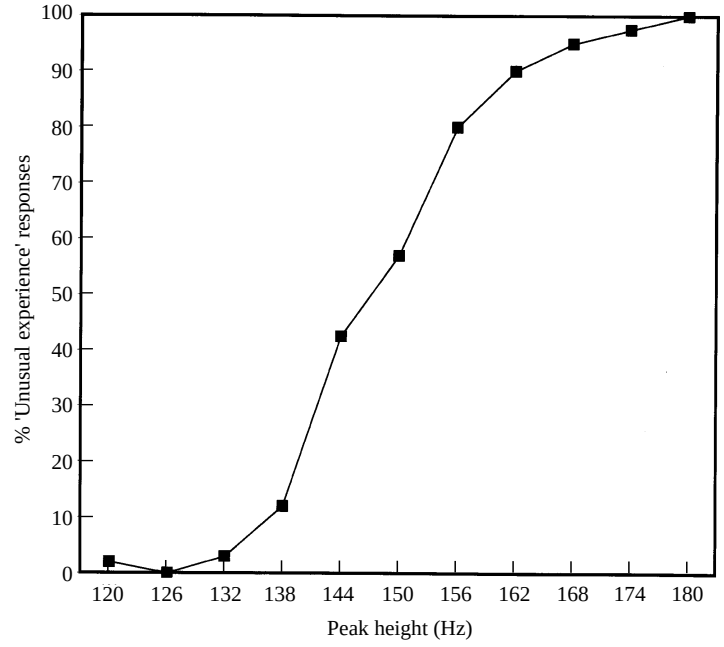

Figure 8. Results of identification task, Experiment 2b.

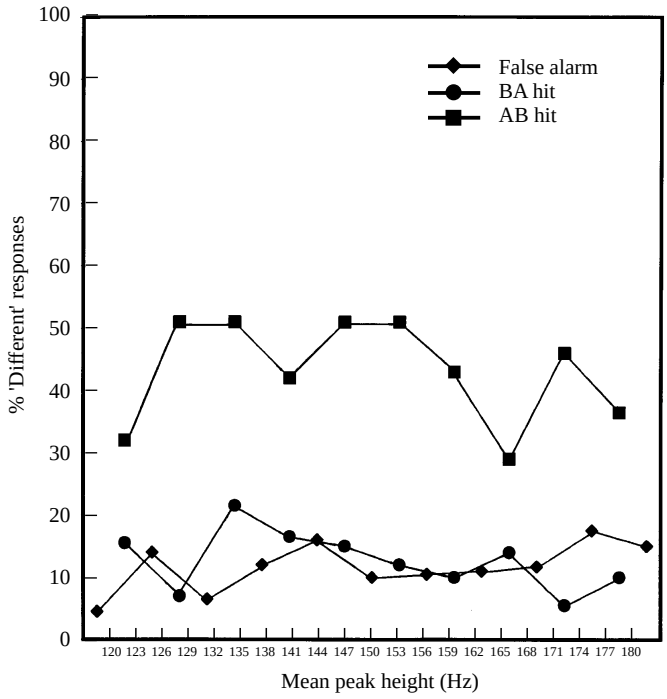

Figure 9. Results of discrimination task, Experiment 2b.

"everyday occurrence" responses at the low end of the peak f_0 continuum and a preponderance of "unusual experience" responses at the high end. There appears to be little difference between these results and the corresponding results of Experiment 2a (Fig. 6). Probit analysis suggests that the boundary between the two categories lies at 148.7 Hz (compared to 144.9 Hz in Experiment 2a). This part of the results of Experiment 2a can be considered replicated. Moreover, we may conclude that the difference between the TR and SL method of stimulus preparation does not affect responses in the identification task.

4.2.2.2. *Discrimination task.* Results of the discrimination task are presented in Fig. 9. Up to a point, we can see similarities between the results of this experiment and those of Experiment 2a: in particular, we see that the false alarm rate gradually increases with increasing peak f_0, and that the percentage of "hits" for the AB pairs is better in the middle of the stimulus continuum than at the ends (and in fact is slightly above the 50% level this time). However, if we subtract the false alarm rate from the AB hit rate (i.e., if we correct for the apparent response bias) we find a discrimination peak at 135 Hz, instead of the 149 Hz predicted by the probit analysis. In other words, these results, while not grossly inconsistent with the general idea of a categorical distinction of some sort between normal and emphatic, are even less compatible with the hypothesis of strict CP than the results of Experiment 2a. Like the results of the identification task, the results of the discrimination task also suggest that the use of the SL method for creating stimuli instead of the TR method does not seriously affect the outcome of these experiments, and allows us to use the SL method in preparing the stimuli for Experiment 3 (described below).

The most striking feature of the results of this experiment has to do with the newly introduced BA discrimination pairs, i.e., the pairs in which the second stimulus has a lower peak than the first. It can be seen from Fig. 10 that there is no difference between the hit rate for these BA pairs and the false alarm rate. That is, it appears that subjects cannot discriminate stimulus pairs presented in BA order, even though they can clearly discriminate the same pairs in AB order. The ANOVA results show a significant presentation order effect between AB and BA [$F(1, 18) = 104.962$, $p < 0.0005$], and show that there is no significant difference between the number of ''different'' responses given to the BA order and the number of false alarms [$F(1, 18) = 1.404$, $p = 0.2514$].

There is no obvious explanation for such an order effect, although it is easy to speculate that it might be related to the findings of ''declination'' experiments (e.g., Pierrehumbert, 1979; Gussenhoven & Rietveld, 1988; Terken, 1991) in which the second accent peak in an utterance is perceived to be equally prominent to the first when it is actually slightly lower in f_0. We return to discuss this in connection with the results of Experiment 4.

5. Experiment 3

One criticism of forced choice identification tasks is that, given any acoustic continuum, subjects will attempt to categorize it. Thus it is difficult to be sure that categorical identification functions result from an intrinsic property of the stimuli, or whether they are simply the consequence of how most subjects approach the identification task itself. However, if the category boundary can be shown to shift in a predicted direction when the stimuli are manipulated in some way, the argument for considering the identification function to be the result of a real perceptual phenomenon rather than test-taking strategies is strengthened.

Although the case for strict CP in Experiments 2a and 2b is not proven, there was nevertheless clear evidence of some sort of categorization in the identification tasks of both experiments. In keeping with our original plan (Section 2.), we therefore decided to investigate the nature of the category boundary by trying to identify factors that would cause it to shift. The most obvious such factor is peak alignment (PA).

Results from the pilot experiment and Experiment 1 suggested that later PA has the effect of increasing the emphasis of an utterance. PA should therefore have an effect on boundary position: we should expect that the boundary between normal and emphatic will be at a lower peak f_0 in stimuli with later PA than it will be for stimuli with earlier PA. This was the hypothesis tested in Experiment 3.

5.1. *Method*

5.1.1. *Stimuli*

We created two sets of stimuli, one with early PA and one with late PA. In order to manipulate PA, as we have already noted, it was necessary to use the SL method rather than the TR method for creating the contours. Comparison of Experiments

2a and 2b suggests that the choice of one method over another had little or no effect on the results of either the identification task or the discrimination task. Since Experiment 3 involved exactly the same kinds of tasks, we felt that the SL method should yield results comparable to both Experiments 2a and 2b.

The 11 stimulus types from Experiment 2b were used as the Early PA set, and a second set of stimuli was created for use as the Late PA set. The same Normal Source utterance of speaker RL's *alarm* sentence, and the same set of f_0 values at each target (see Appendix), were used for creating the Late PA set. The only difference was that the timing of the four targets that defined the accentual pitch excursion was delayed by 60 ms. This had the effect of moving the whole pitch accent later, while maintaining the slopes that lead up to and away from the Peak target. The peak in the late PA continuum was aligned with the very end of the stressed [a:] of *alarm*.

5.1.2. *Procedures*

In general the experimental procedures were the same as for Experiment 2a and 2b. There were 13 subjects, spread over 2 experimental sessions.

In the identification task, the procedures here were essentially identical to Experiments 2a and 2b. The stimulus set for the practice session consisted of five tokens of Step 1 (120 Hz peak) from the Early peak continuum, and five tokens of step 11 (180 Hz peak) from the Late peak continuum. These two stimuli were assumed to represent the two extremes of the entire stimulus set. As before, the practice stimuli were randomized.

In the discrimination task, the stimuli from the Early and Late continua were kept separate in creating discrimination pairs. The four continua of pair types AB, BA, AA, and BB which were used in Experiment 2b were used again in Experiment 3 for the Early PA condition. An analogous set of pairs was created for the Late PA condition. There were thus 80 different pairs altogether. Three tokens of each pair were randomized and recorded onto a tape with a one-second intrastimulus interval, and a 2.5-second inter-pair interval including a half second warning tone. The inter-pair interval was reduced in this experiment in order to cut down the length of the experimental sessions. A practice session was given before the main part of the discrimination task began, with a representative sample of stimuli from each of the 8 continua.

In other respects the experimental procedures were identical to those of Experiments 2a and 2b.

5.2. *Results*

5.2.1. *Identification task*

Results of the identification task are shown in Fig. 10. There does appear to be a boundary shifting effect as predicted. The Late PA response curve shows an abrupt rise in "unusual experience" responses from step 3, or 132 Hz, whereas in the Early PA curve the abrupt shift does not begin until step 5, or 144 Hz. This is consistent with the idea that Late PA is intrinsically more emphatic, and has the effect of shifting the category boundary in a continuum of peak f_0.

In order to test the significance of this apparent effect, we used probit analysis on

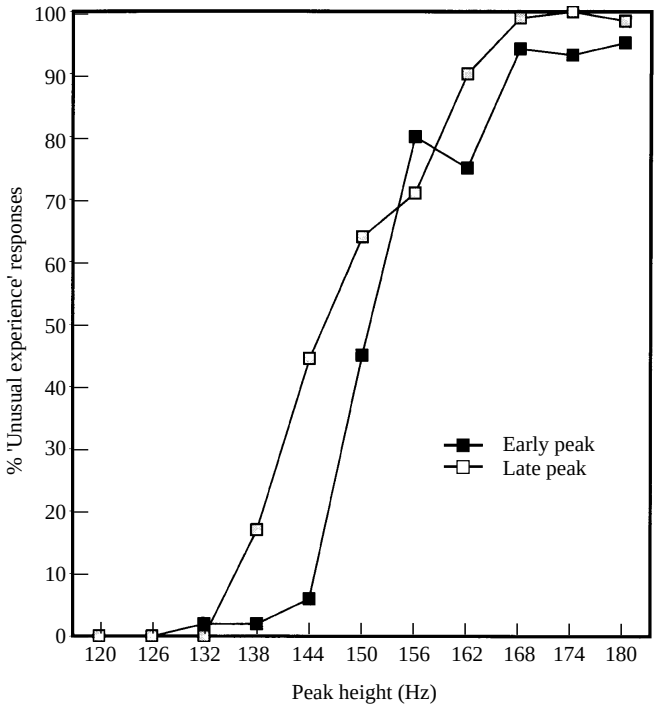

Figure 10. Results of identification task, Experiment 3.

the 13 individual subjects' identification data to determine their Early PA boundary and Late PA boundary. In all 13 subjects the Late PA boundary occurred at a lower peak f_0 than the Early PA boundary (mean Early PA boundary 153.8 Hz, mean Late PA boundary 148.1 Hz). This difference in boundary values was found to be significant by a one-way ANOVA [$F(1, 24) = 10.753$, $p < 0.005$].

5.2.2. *Discrimination task*

Results of the discrimination task are shown in Fig. 11. As in Experiment 2b, the "hit" rates for the BA pairs were not significantly different from the false alarm rates and the BA results are omitted from the figure. As in Experiments 2a and 2b, the "hit" rate for the AB pairs was significantly different from the false alarm rate and showed a modest peak somewhere in the middle of the continuum. This peak, also as in Experiments 2a and 2b, seemed less convincing under closer examination than it appeared at first.

For the Early PA stimuli, considering only the discriminable AB pairs, there is a slight peak in discrimination at 153 Hz, which is where the identification function would predict a peak. However, 153 Hz is also the place where the number of false alarms is the highest. If we correct the plot of hits by subtracting the false alarm rate, we find no obvious peak at all. In the Late PA stimuli, there is a more noticeable peak, reaching 60 percent "Different" judgments at 141 Hz, which remains as the discrimination peak when the rate of false alarms is taken into account. Unfortunately, 141 Hz is one step earlier than the mean Late PA identification boundary of 148 Hz.

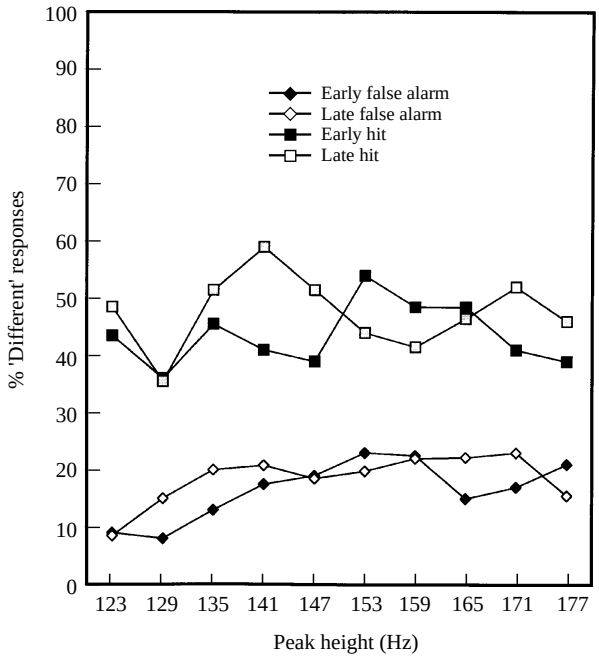

Figure 11. Results of discrimination task, Experiment 3.

5.3. *Discussion*

Once again, the evidence for CP in the strict sense is weak, but once again the identification judgements show a clear S-shape. Moreover, as predicted, the manipulation of PA shifts the boundary without affecting the abruptness of the shift from one judgement category to the other. This strengthens the conclusion from the preceding experiments that we are not dealing with CP of the sort that is found in perception of segmental phonemes. At the same time, it does appear that listeners are predisposed to interpret accents as categorically either normal or emphatic, rather than consistently interpreting fine differences of pitch range in terms of fine differences in degree of emphasis.

6. Experiment 4

6.1. *Introduction*

The results of the discrimination tasks in Experiments 2a, 2b, and 3 suggested that our stimulus pairs were fairly close to the threshold of perceptibility. It is possible that this would explain why the results of these experiments were, from the point of view of the classical CP paradigm, fairly inconclusive. One way to proceed might therefore have been to re-run the experiments with larger differences between the members of the stimulus pairs. However, as we noted earlier, a pilot experiment suggested that increasing the gap between the stimuli from one step (6 Hz) on the peak f_0 continuum to two steps (12 Hz) would have resulted in stimulus pairs that were fairly generally discriminable without regard to the location of any putative

category boundaries. The very perceptibility of small differences within putative categories obviously makes the idea of CP rather difficult to maintain.

More importantly, it makes the classical CP *methodology* difficult to apply. The problem is that listeners, even though they seem to *interpret* the pitch range on a one-accent utterance categorically as either normal or emphatic, nevertheless remain equally able (or in some cases, unable) to *perceive* fine differences of detail between the pitch range of one accent and that of another, whether within or across categories. This state of affairs cannot be accommodated within classical CP methodology, which relies on the virtually complete inability of subjects to perceive fine differences of detail within categories. That inability is what makes the same-different judgement task meaningful, and that inability is precisely what, in the case of pitch range, we cannot count on. Consequently, we felt it would probably be pointless to try to apply classical CP methodology any further.

However, even if subjects are always going to be inclined to respond "Different" to pitch range stimulus pairs in a classical CP discrimination task—provided the difference is above threshold—we reasoned that they might nevertheless be able to make reliable judgements that one stimulus pair is *more different* than another. This suggested a different approach to looking for peaks of discriminability. Specifically, we presented subjects with a considerable variety of stimulus pairs, and asked them to rate, on a ten-point scale, how different from each other the two members of the pairs seemed. We predicted that, if there is a peak of discriminability across the putative category boundary, it should be reflected in *higher difference ratings* for stimulus pairs that straddle the boundary than for stimulus pairs that lie within one category. As with classical CP methodology, in other words, we were still looking for a peak in the discrimination results that would correspond to the boundary in the identification results; the only difference was that the peak we were looking for was not a peak in absolute same-different judgements, but rather a peak in the degree of difference that subjects detected between the members of a stimulus pair.

6.2. *Method*

In order to keep the experimental sessions to a manageable length we dispensed with the identification task and used only the new type of discrimination task. Since the stimulus pairs were constructed from stimuli used in Experiments 2b and 3, we decided to use the results of the identification tasks of those experiments as an indication of the boundary location on the peak f_0 continuum.

6.2.1. *Stimuli*

The stimulus pairs were constructed from the same 11 Early PA stimuli used in Experiments 2b and 3. We constructed all possible pairs of stimuli that did not differ at all (the equivalent of the "AA" and "BB" stimulus pairs in Experiment 2b, but we used only one set), and all possible pairs that differed by one, two, and three steps along the peak f_0 continuum. For example, the two-step pairs were step $1 + 3$, step $2 + 4$, and so on up to step $9 + 11$; the three-step pairs were step $1 + 4$, step $2 + 5$, and so on up to step $8 + 11$. All the pairs whose members differed from one another were constructed in both orders ("AB" and "BA" pairs). This resulted in a total of $11 + 2 \times (10 + 9 + 8) = 65$ stimulus pairs.

The 65 stimuli pairs were randomized in three different ways to create three blocks, each containing all 65 stimuli. These 195 stimuli were recorded onto a tape with 1 s intrastimulus interval and a 2.5 s interstimulus interval containing a 0.5 s warning tone. A second tape was prepared in the same way, except that within each block, the order of the 65 stimuli was reversed.

6.2.2. *Procedures*

A practice session of 18 stimulus pairs was presented before the main part of each session, in order to make subjects aware of the range of similarity or difference. The 18 pairs in the practice session consisted of 4 ''same'' pairs, 4 three-step AB pairs, and two examples of all other pair types, giving a representative range of peak f_0. The first five stimuli in the practice were three ''same'' pairs followed by two three step AB pairs. The 13 stimuli that followed were randomized. Subjects were given instructions that the practice session would contain the full range of similarity or difference, and that they should base their use of the rating scale upon what they heard in the practice.

The screen display consisted of a rating scale with ten boxes. The poles of the scale were labelled ''Least Different/Almost Identical'' at the low end, and ''Most Different'' at the high end. An arrow over the boxes indicated the direction of increasing difference. Subjects were asked to rate ''how different each pair sounds'' on the ten-point scale.

Two groups of subjects took part in the experiment. Each group heard both stimulus tapes on separate occasions, roughly four days apart. The first group, consisting of 6 subjects, heard tape 1 at the first session and tape 2 at the second session. The second group, consisting of 5 subjects, heard the tapes in the reversed order. In total, each subject gave 6 responses to each of the 65 stimulus pairs.

A short break was given half way through each block of 65, as well as at the end of each block. Each session lasted about 30 minutes.

6.3. *Results and discussion*

Results, for the AB pairs and BA pairs separately, are shown in Fig. 12. It can be seen that there is a steady increase in the difference judgements as the size of the gap in the stimulus pair increases: in general, three-step stimulus pairs are judged more different than two-step stimulus pairs, which in turn are judged more different than one-step stimulus pairs. That is, stimulus pairs that are objectively more different are in general rated more different. This suggests that the task is a valid approach to probing listeners' percepts of differences between contours. Moreover, as in Experiments 2b and 3, there was no difference between judgements of the one-step BA pairs and the identical pairs, whereas there was a difference between the judgements of the one-step AB pairs and the identical pairs. This consistency between the classical CP discrimination task and the present difference-rating task further suggests the validity of our approach.

However, the results are not at all as predicted. There are no clear peaks in the discrimination functions in the vicinity of the boundaries, and in fact some of the plots show an apparent *valley* at about the place where we would have predicted a

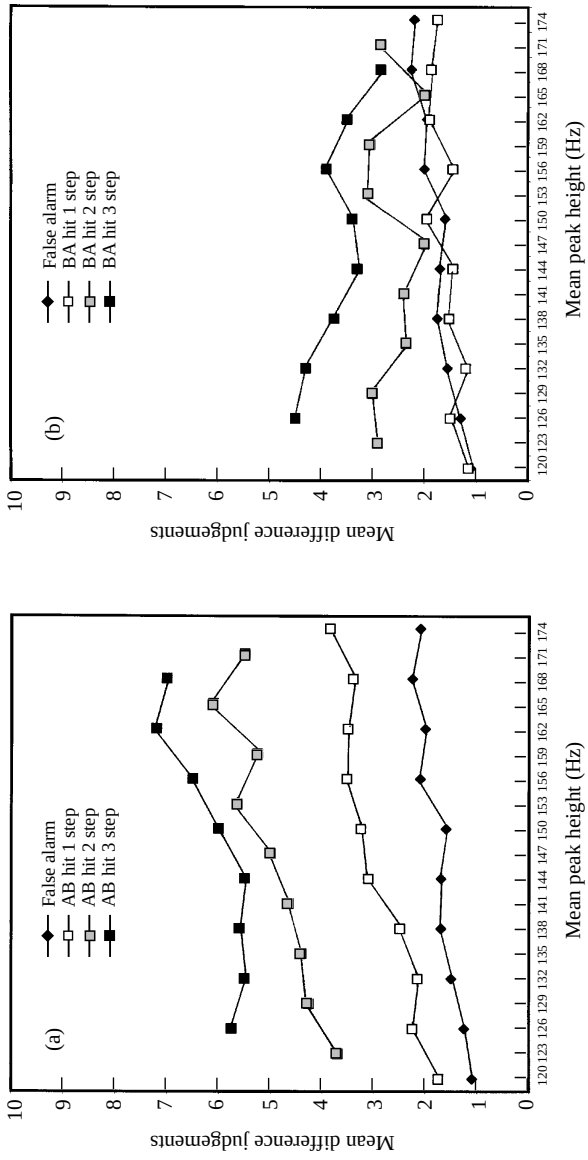

Figure 12. Overall results of Experiment 4, (a) pairs in AB order, (b) pairs in BA order.

peak. We decided not to pursue the search for discrimination peaks any further, and accept that the apparent abrupt shifts in the identification functions for our stimuli are not comparable to the boundaries between phonological categories. Implications of this for the interpretation of pitch range are discussed in Section 7 below.

While the results do not show the discriminability differences that we predicted, they bring to light an entirely unsuspected type of difference for which we have no explanation at present. This is the difference between the results for the AB pairs and the BA pairs, which can be seen in Fig. 12. First, there is the simple difference between the one-step AB and one-step BA pairs, namely that the former are discriminably different from identical pairs and the latter are not. This replicates the findings of Experiments 2b and 3 and makes it appear that this order-of-presentation effect is real and needs explanation. As noted above, it seems plausible to relate this effect to the effect of declination on the perception of relative prominence of accents in short utterances.

Second and considerably more puzzling, the *size* of this order-of-presentation effect appears to depend very strongly on the place of the stimuli on the peak f_0 continuum. This can also be seen in Fig. 12. The AB pairs [Fig. 12(a)] show a general trend first noted in Experiment 2a for the false alarm data: the higher the peak f_0, the more different the members of the pair sound. But the BA pairs [Fig. 12(b)] show the opposite effect; the *lower* the peak f_0, the more different the members of the pair sound. Comparing the actual judgement values between Fig. 12a and 12b, we can see that the AB and BA judgement curves lie close together at the low end of the peak f_0 continuum, but diverge at the high end. Put differently, comparison of Fig. 12(a) and 12(b) reveals a large order-of-presentation effect favoring AB order over BA order, at high peak f_0, but only a very small order-of-presentation effect at relatively low peak f_0. Why the order effect should change like this is, to us, a complete mystery. However, we note that a few other studies (e.g., Repp, Healy & Crowder, 1979; Schiefer & Batliner, 1991; Verhoeven, 1991) have reported finding variable order-of-presentation effects under certain circumstances. Traditionally, order effects are treated as little more than a methodological nuisance that must be ''controlled for'' in designing experiments. On the basis of our discovery here, we feel that systematic investigation would be extremely valuable.

7. Summary and conclusion

The work reported in this paper was motivated by theoretical speculations and preliminary empirical findings that suggested the possible existence of a category boundary between ''normal'' and ''emphatic'' accents along the scale of increasing pitch excursion. The evidence pointing to such a boundary included the following: (i) the existence of clearly distinct interpretations for short utterances, when the utterances are presented with intonation contours that are identical except for pitch range (Hirschberg & Ward, 1992); (ii) the existence of two apparently distinct modes of interpreting the relative height of adjacent accents in an utterance, one for accents with moderate pitch range and one for accents with wide pitch range (Gussenhoven & Rietveld, 1988; Ladd *et al.,* 1994); and (iii) the preliminary finding that judgements of the degree of emphasis conveyed by a single-accent utterance

show a relatively abrupt shift from ''normal'' to ''emphatic'' judgements—i.e. S-shaped response curves—as a function of gradually increasing pitch excursion on the accent contour (Morton, 1993).

On the basis of this evidence, we tested whether the putative boundary between normal and emphatic accents in single-accent utterances is a boundary of the sort found between segmental phonemes in a language, e.g., the boundary between /p/ and /t/ or between /t/ and /d/ in English. Specifically, we looked for evidence of the ''categorical perception'' that often characterizes such phoneme boundaries. Our data offer little support for a finding of classical categorical perception. Under a variety of conditions we do find evidence of S-shaped response curves, but only one experiment yielded a discrimination function that even approximates a discrimination peak at the category boundary. That is, it appears that listeners are able to discriminate fairly small distinctions of pitch range in pairs of single-accent utterances, irrespective of where the stimulus pairs are located on a continuum from very unemphatic to very emphatic.

However, it is worth emphasising that we do find S-shaped response curves under a variety of conditions. That is, listeners' judgements of the pragmatic force of accents can shift fairly abruptly from ''normal'' to ''emphatic'' as a function of increasing pitch range, even though these shifts are apparently not accompanied by the heightened discriminability characteristic of classical categorical perception. While we cannot therefore legitimately speak of pitch range distinctions being categorically *perceived,* it may be useful to think of them as being categorically *interpreted.* That is, it does not seem unreasonable to suggest, as we did in Section 5.3., that listeners are predisposed to interpret accents or utterances as being categorically either ''normal'' or ''emphatic''. A variety of acoustic and pragmatic parameters play a role in this decision, including pitch range, voice quality, lexical content, discourse background, relationship between the speaker and listener, and so on. Any or all of these parameters may be continuously variable, and the continuous variability may be directly perceptible as such, and there is thus no true categorical perception. Yet the interpretation computed on the basis of all the input parameters nevertheless normally falls unambiguously into one category or the other.

Nevertheless, there remains stubborn evidence of puzzling perceptual effects of some sort. The ''phonological'' explanation proposed by Ladd *et al.* (1994) for the original Gussenhoven–Rietveld effect was based on the idea that there might be a true phonological boundary between normal and emphatic accents. If this possibility is ruled out—as it appears to be by the present study—then the Gussenhoven–Rietveld effect remains unexplained. Moreover, if we rule out the phonological explanation, it appears more likely that some genuine perceptual or psychophysical effect is at work in Gussenhoven and Rietveld's findings (and their replication and extension by Ladd *et al.*). This likelihood is increased by the findings of Experiment 4 of the present study, which must for the moment remain entirely unexplained. Experiment 4 shows very clearly that the perceived similarity of two single-accent utterances differing only in pitch excursion depends quite substantially on the order in which the two are presented, and, more importantly, that this order effect is not constant throughout the continuum of pitch range. One could imagine ways in which this finding might be used to construct an explanation for the Gussenhoven–Rietveld effect.

However, until the underlying order effect is understood there seems little point in trying to explain relatively specific manifestations such as the Gussenhoven–Rietveld effect. In effect, the chain of investigation that began with Gussenhoven and Rietveld's accidental discovery has so far been much more successful in raising questions than in finding answers. We hope that the order effect discovered here may stimulate research that will provide answers to fundamental questions about the perception of relative prominence and emphasis in spoken language—answers to questions that, so far, we are only dimly aware of the need to ask.

The research reported here was supported by the UK Economic and Social Research Council (ESRC) through grant no. R000 22 1111 to Edinburgh University. This support is gratefully acknowledged. We also thank Irene Macleod, Cedric Macmartin, Mike Morony, and Steve Isard for assistance with programming and data analysis. During part of the period of the grant the first author was a visiting researcher at the Institute for Perception Research (IPO) in Eindhoven, The Netherlands, an opportunity for which he is extremely grateful. We thank several IPO colleagues, particularly Dik Hermes, Willem Rump, Jacques Terken, Aad Houtsma, and Huib de Ridder, for useful discussion.

References

Bolinger, D. (1961) *Generality, gradience, and the all-or-none.* The Hague: Mouton

Crystal, D. (1969) *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press

Finney, D. J. (1971) *Probit analysis.* Cambridge: Cambridge University Press

Glasberg, B. R. & Moore, Brian (1990) Derivation of auditory filter shapes from notched-noise data, *Hearing Research,* **47**, 103–138

Gussenhoven, C. & Rietveld, T. (1988) Fundamental frequency declination in Dutch: testing three hypotheses, *Journal of Phonetics,* **16**, 355–369

Harnad, S., editor, (1987) *Categorical perception: the groundwork of cognition.* Cambridge: Cambridge University Press

Hermes, D. & van Gestel, J. (1991) The frequency scale of speech intonation, *Journal of the Acoustical Society of America,* **90**, 97–102

Hirschberg, J. & Ward, G. (1992) The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English, *Journal of Phonetics,* **20**, 241–251

Kohler, K. J. (1987) Categorical pitch perception. In *Proceedings of the 11th International congress of the phonetic sciences,* Tallinn, **5,** pp. 331–333

Ladd, D. R. (1983) Phonological features of intonational peaks, *Language,* **59**, 721–759

Ladd, D. R. (1990) Metrical representation of pitch register. In *Papers in laboratory phonology I,* (J. Kingston & M. Beckman, editors), pp. 35–57. Cambridge: Cambridge University Press

Ladd, D. R. (1994) Constraints on the gradient variability of pitch range (or) Pitch Level 4 Lives! In *Papers in laboratory phonology III* (P. Keating, editor), pp. 43–63. Cambridge: Cambridge University Press

Ladd, D. R. (1996). *Intonational phonology.* Cambridge: Cambridge University Press

Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G. & Scherer, K. R. (1985). Evidence for the independent function of intonation contour, pitch range, and voice quality, *Journal of the Acoustical Society of America,* **78**, 435–444

Ladd, D. R., Verhoeven, J. & Jacobs, K. (1994) Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects, *Journal of Phonetics,* **22**, 87–99

Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957) The discrimination of speech sounds within and across phoneme boundaries, *Journal of Experimental Psychology,* **61**, 379–388

Liberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In *Language sound structure* (M. Aronoff & R. Oerhle, editors), pp. 157–233. Cambridge, MA: MIT Press

Menn, L. & Boyce, S. (1982) Fundamental frequency and discourse structure, *Language and Speech,* **25**, 341–383

Morton, R. (1993) *On the structure of pitch range in intonation: gradient or categorical?* Honours Dissertation, University of Edinburgh

O'Connor, J. D. & Arnold, G. F. (1973) *Intonation of colloquial English* (2nd ed.). London: Longman

Pierrehumbert, J. (1979) The perception of fundamental frequency declination, *Journal of the Acoustical Society of America,* **66**, 363–369
Pierrehumbert, J. (1980) *The phonology and phonetics of English intonation.* PhD Dissertation, MIT
Pike, K. L. (1945) *The Intonation of American English.* Ann Arbor: University of Michigan Press
Repp, B. (1984) Categorical perception: issues, methods, findings. In *Speech and Language: advances in basic research and practice* (N. J. Lass, editor), **10**, New York: Academic Press
Repp, B. H., Healy, A. F. & Crowder, R. G. (1979) Categories and context in the perception of isolated steady-state vowels, *Journal of Experimental Psychology: human perception and performance,* **5,** 129–145
Schiefer, L. & Batliner, A. (1991) A ramble round the order effect, *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München,* **29**, 125–180
Shriberg, E. E., Ladd, D. R., Terken, J. & Stolcke, A. (1996) Modeling pitch range variation within and across speakers; predicting f_0 targets when ''speaking up''. In *Proceedings of International Conference on Spoken Language Processing,* Philadelphia, supplement pp. 1–4
Silverman, K. (1987) *The structure and processing of fundamental frequency contours.* PhD thesis, Cambridge University
Terken, J. (1991) Fundamental frequency and perceived prominence of accented syllables, *Journal of the Acoustical Society of America,* **89,** 1768–1776
't Hart, J., Collier, R. & Cohen, A. (1990) *A perceptual study of intonation: an experimental-phonetic approach.* Cambridge: Cambridge University Press
Thorsen, N. (1980) A study of the perception of sentence intonation—evidence from Danish, *Journal of the Acoustical Society of America,* **67**, 1014–1030
Trager, G. L. & Smith, H. L. (1951) *An outline of English structure.* Norman, OK: Battenburg. (Reprinted in 1957 by American Council of Learned Societies, Washington)
Verhoeven, J. (1991) *Perceptial aspects of Dutch intonation.* PhD Thesis, University of Edinburgh
Wells, R. (1945) The pitch phonemes of English, *Language,* **21,** 27–40

Appendix: details of f_0 contour modification

1. *Establishment of standard contours*

On the basis of the full set of recordings (cf. Section 2.1.), we established values for idealized ''normal'' and ''emphatic'' pitch contours for each of the four speakers. These idealized values are referred to in what follows as ''standard target values''. There are four target points in each utterance. These are: Abs (the first reliable f_0 point or absolute onset of voicing); Onset (the vowel onset of the accented syllable); Peak (the highest f_0 value in the contour); and Endpoint (the last reliable f_0 point).

Values for the first three standard targets were defined as the means of the measurements from 9 utterances (3 repetitions $\times$ 3 sentences); we computed separate normal and emphatic standards for these three targets. Endpoint, by contrast, was assumed not to vary with pitch range (cf. Menn & Boyce, 1982; Liberman & Pierrehumbert, 1984; Shriberg, Ladd, Terken & Stolcke, 1996), and was therefore calculated as the mean final f_0 of all 18 utterances. This mean was then used as the standard Endpoint target for both normal and emphatic utterances.

On the basis of these standard target values, we created a number of *source contours,* which were the contours actually used as the basis for the rescaling of pitch range in preparing the experimental stimuli. For each set of stimuli, the source contour was imposed on the source utterances, and the pitch range was then modified to create the full stimulus set.

When we used the Straight Line (SL) method of rescaling (see Sections 2.2. and 5.1.1.), it was necessary to add a few extra targets to the four basic ones in order to keep the shape of the resulting contour approximately congruent with the contours prepared by the Total Rescaling (TR) method. In all cases a ''Pre-Peak'' and an ''Elbow'' were added; the former serves to simulate the rounding off of the peak of

the accentual contour, while the latter indicates the location between the Peak and the Endpoint where the rapid drop in f_0 flattens out. In addition, two Pre-End targets were added at the end of the word *went* and the start of the word *off* in the Alarm text, because the f_0 contour did not follow smoothly across the silent gap between these two words (many British speakers have a fully articulated and voiceless [t] in this context).

2. *Pitch range modification model*

As noted in the text (Section 2.2.), a quantitative model of pitch range is required for rescaling contours. For this purpose, we adopted the model used in work at the University of Giessen by Ladd, Silverman, Tolkmitt, Bergmann & Scherer, 1985 and referred to here as the Giessen model. The Giessen model relates f_0 values in one pitch range to f_0 values in another pitch range by the following formula, where Fr is the speaker ''Floor'' value:

$$\log [f_0(\text{range } 2)/\text{Fr}] = \text{R} \log [f_0(\text{range } 1)/\text{Fr}]$$

Note that this can be used with either the TR or SL methods of generating modified contours for use in stimuli (see Section 2.2. for more detail on these two methods). In the SL method, the Giessen formula is applied only to the target points; in the TR method, it is applied to all points.

Before using the Giessen model, we tested it against normal and emphatic standard values. If the model is valid, it should be possible to generate the emphatic standard targets from the normal standard targets and vice-versa. Specifically, it should be possible to calculate a value for *R* on the basis of the standard target values for the normal and emphatic Peaks, and then use that value of *R* to predict the emphatic standard targets values of Abs and Onset from the normal standard target values, or vice versa. In fact, in every case the model's predictions for Abs and Onset—the ''Prehead'' values—were slightly but consistently in error: if we scale up the normal Prehead on the basis of the *R* derived from the Peaks, we predict emphatic Prehead values that are higher than those actually observed; and of course conversely, if we scale down emphatic Prehead values on the basis of the Peak, we predict lower normal Prehead values than we observe.

Despite this error, we used the Giessen model anyway, for a number of reasons. First and most importantly, our goal was not to model production data accurately but simply to create natural sounding stimuli for our perception experiments. Second, we could compensate for the error by dividing the contour in half at the Peak, applying the Giessen model to the first half and the second half of the contour separately using an ad hoc value for Fr. Third, we wanted in any case to manipulate the scaling of the Prehead as an independent variable in our experiments, which actually required us to split the contour in half as just described. As noted in the text, these manipulations of the prehead led to inconclusive results and are not reported in the paper.