Integrality in the perception of tongue root position and voice quality in vowels

John Kingston, Neil A. Macmillan, Laura Walsh Dickey, Rachel Thorburn, and Christine Bartels

Citation: The Journal of the Acoustical Society of America **101**, 1696 (1997); doi: 10.1121/1.418179
View online: https://doi.org/10.1121/1.418179
View Table of Contents: https://asa.scitation.org/toc/jas/101/3
Published by the Acoustical Society of America

Integrality in the perception of tongue root position and voice quality in vowels

John Kingston
Linguistics Department, South College, University of Massachusetts, Amherst, Massachusetts 01003

Neil A. Macmillan
Psychology Department, Brooklyn College of the City University of New York, Brooklyn, New York 11210

Laura Walsh Dickey,[a] Rachel Thorburn, and Christine Bartels
Linguistics Department, South College, University of Massachusetts, Amherst, Massachusetts 01003

(Received 5 December 1995; revised 25 June 1996; accepted 23 October 1996)

In English and a large number of African and Southeast Asian languages, voice quality along a tense–lax dimension covaries with advancement of the tongue root in vowels: a laxer voice quality co-occurs with a more advanced tongue root. As laxing the voice increases energy in the first harmonic relative to higher ones and advancing the tongue root lowers F_1, the acoustic consequences of these two articulations may integrate perceptually into a higher-level perceptual property, here called spectral ''flatness.'' Two Garner-paradigm experiments evaluated this interaction across nearly the entire range of tense–lax voice qualities and a narrow range of F_1 values. The acoustic consequences of laxness and advanced tongue root integrated into spectral flatness for tenser and laxer but not for intermediate voice qualities. Detection-theoretic models developed in earlier work proved highly successful in representing the perceptual interaction between these dimensions. © *1997 Acoustical Society of America.* [S0001-4966(97)03203-7]

PACS numbers: 43.71.An, 43.71.Es, 43.71.Hw, 43.70.Fq [RAF]

INTRODUCTION

That minimally contrasting speech sounds differ along multiple dimensions challenges both phoneticians and psychophysicists. Phoneticians are challenged to explain why some patterns of multiple differences recur across languages but others do not. Psychophysicists are challenged to explain how the array of differences between sounds influences internal representations or processes in the listener. However, neither challenge can be met in a sufficiently general way alone. General psychophysical models of how an external stimulus relates to an internal category are one way of explaining why some arrays of differences recur across languages, and patterns that recur across languages reveal some of those psychophysical processes and representations which are likely to arise naturally in response to stimuli.

The specific questions addressed by this paper are: (1) do the many differences between minimally contrasting speech sounds interact perceptually?; (2) if they do, how do they interact? and finally; (3) does the way in which they interact contribute to explaining why these patterns recur across languages? Answers to all these questions require explicit and general psychophysical models of these interactions.

The phonetic and psychophysical challenges are taken up in turn below.

A. The phonetic challenge

Previous work examining perceptual interactions among arrays of differences that recur across languages has examined the multiple differences between vowels contrasting for height (Kingston, 1991; Hoemeke and Diehl, 1994; Kingston and Macmillan, 1995) and intervocalic stops contrasting for [voice] (Parker *et al.*, 1986; Kluender *et al.*, 1988; Diehl and Kingston, 1991; Kingston and Diehl, 1995; Diehl *et al.*, submitted). Here, the focus is on the interaction between a source and a filter property of vowels, specifically, between voice quality and the position of the tongue root.

Vowels articulated with advanced tongue root (ATR) are often also produced with a lax or breathy voice quality, whereas vowels articulated with retracted tongue root (RTR) are often produced with a tense or creaky voice quality. This covariation of voice quality with the position of the tongue root is observed in many but not all[1] languages in East and West Africa in which vowels harmonize for ATR/RTR [Hall *et al.*, 1974; Lindau, 1975, 1978, 1979; Jackson, 1988; see Denning, 1989, for an extensive literature review and Jacobsen (1978, 1980) for a more specific discussion of such facts in DhoLuo and other Nilotic languages].

An essentially similar pattern of covariation can be also be observed in the historical development of many Mon–Khmer languages of Southeast Asia, in which higher vowels have developed diachronically in syllables which originally had breathy or lax voice, while lower vowels developed in originally tense (or modal) voiced syllables (see Huffman, 1976 for a review). As raising the tongue body frequently entails advancing the tongue root as well (Perkell, 1969; Jackson, 1988), the Mon–Khmer pattern is fundamentally similar to that observed in the African languages.

Finally, tense vowels in American English have laxer voice qualities than their lax counterparts (Bloedel, 1994). Differences in tongue position between tense and lax vowels

[a]Now at the Max Planck Institut für Psycholinguistik, PB 130, Nijmegen, NL 6500 AH, The Netherlands.

resemble those between vowels contrasting for (ATR): lax vowels in English are produced with more retracted tongue roots than tense vowels (Perkell, 1969; Jackson, 1988), and Baer *et al.* (1988) report greater hyoglossus contraction, which will retract the tongue root, in lax than tense vowels. Thus, voice quality and tongue root position covary in a similar way in this language as in Southeast Asian and African languages described above.

There are two strong and thus easily falsified ways to explain these patterns of covariation between the position of the tongue root or body and the tension of the vocal folds. Voice quality could vary with tongue root position because one articulation is linked physiologically to the other. Alternatively, these articulations could be covaried by speakers because their acoustic consequences enhance one another's perceptual effects. (A third, far less easily falsified, and thus inherently weaker kind of explanation will be laid out after these two stronger kinds are developed.) These two explanations are not mutually exclusive: speakers may exaggerate the covariation between physiologically linked articulations in order to enhance the perceptual effects of each's acoustic consequences. They are laid out separately below to show how each might be falsified.

One way in which voice quality may depend physiologically on tongue root position is if the aryepiglottal ligament and membrane, which connect the tongue root to the arytenoid cartilages via the epiglottis, cause the arytenoids to slide forward slightly and/or rock slightly apart, slackening or separating the vocal folds enough to lax the voice, when the tongue's root is advanced or its body raised.[2] However, vocal fold tension cannot always depend physiologically on tongue root or body position because it is independently controlled in some languages; for example, in Dinka, a Nilotic language of Kenya, vowels may contrast independently for tense versus lax voice quality and advanced versus retracted tongue root position (Denning, 1989).

The perceptual explanation for covariation of voice quality and tongue root position is consistent with this independent control. The perceptual explanation is most compeling if the covarying articulations' acoustic correlates are similar enough psychoacoustically to integrate into higher-level perceptual properties. In this case, the contrast is enhanced because both articulations' correlates evoke the same psychoacoustic property, not simply because the minimally contrasting phonemes differ in more than one way. Laxing the voice increases energy in the first harmonic at the expense of higher ones and advancing the tongue root lowers F_1. Tense voice and a retracted tongue root have the opposite acoustic effects. The two articulations may thus be deliberately covaried in order to depress or elevate the energy concentration in the vowel's spectrum. This overall effect on energy distribution in the vowel's spectrum will be referred to henceforth as spectral ''flatness.'' [3] This perceptual explanation also extends readily to the covariation of lax or breathy voice with higher tongue positions in the Mon–Khmer languages and tenser vowel qualities in English because higher or tenser lingual articulations also lower F_1.

Independent contrast of voice quality and tongue root position like that observed in Dinka is not, furthermore, a problem for the perceptual explanation, which requires that the two articulations be independently controlled. Such control allows speakers to use the articulations as independent contrasts or to combine them in order to enhance a single contrast.

The most direct means of falsifying the perceptual explanation is to show that the perceptual integration of the acoustic consequences of the two articulations simply does not occur, that they instead remain perceptually separate.

In recent work using part of the paradigm used here, Li and Pastore (1995) have shown that whereas two source properties, F_0 and spectral tilt, of vowel-like stimuli do integrate, the source property spectral tilt remains perceptually separate from a filter property, namely the number of peaks in the filter spectrum. Li and Pastore go on to argue that source and filter properties should in principle not integrate because they convey different kinds of information; according to Li and Pastore, source properties convey speaker identity and other paralinguistic information, whereas filter properties convey the linguistic content of the message. This functional explanation for the separability that Li and Pastore found between source and filter properties is clearly wrong, e.g., the source property F_0 conveys linguistic contrasts of tone and intonation. Moreover, voice quality is contrastive in many languages of Africa and Southeast Asia. Nonetheless, Li and Pastore's psychoacoustic claim may still be correct. Perhaps, the common independence of source and filter variation—one can sing the same vowel at different pitches or different vowels at the same pitch—is sufficient for listeners to keep their acoustic consequences separate.

Fowler (1996) argues for source–filter separability on the grounds that their articulatory origins are independent. She argues, specifically, that because the F_0 perturbations caused by vowel (height) or consonant (voice) contrasts are articulatorily independent of the control of F_0 to convey tone or intonation contrasts, these effects on this source property can all be kept perceptually separate.

The demonstration below that the acoustic consequences of varying tongue root position and voice quality do after all integrate perceptually is accordingly of some general interest for models of speech perception, and disconfirms the claims of both Li and Pastore and Fowler that source and filter properties should remain perceptually separate.

Consider now the third, less easily falsified explanation for covariation between articulations. Any combination of multiple difference articulations could make a minimal contrast easier to perceive, on the principle that more than one difference makes a contrast easier to detect reliably than just a single difference. And it is also possible that listeners can readily learn an association between differences from the simple fact of their covariation, and that those differences need neither cohere psychoacoustically nor share a common articulatory origin. Their covariation may instead be quite arbitrary. For discussion of this kind of systematic if arbitrary covariation, see Ohala (1981).[4]

The difficulty with this kind of explanation is its weakness; if even arbitrary covariation is readily learned, then it is very difficult to falsify such a hypothesis. A relatively weak means of falsification would be to show that languages differ

far less from one another in what patterns of covariation they allow than a theory allowing arbitrary covariation would predict. If instead properties which are not acoustically similar covary in languages just as often as acoustically similar ones do, then this weaker sort of explanation would be sufficient to explain the facts. At present, too little is known and there is substantial dispute regarding the extent to which languages differ from one another in this regard (see Kingston and Diehl, 1994, 1995; and Nearey, 1995, for discussion and opposing points of view). This kind of explanation is more surely falsified by the fact that integration occurs with nonspeech analogues (Kingston and Diehl, 1995; Diehl *et al.*, submitted), for which no arbitrary association could have been learned; see the papers cited for fuller discussion.

B. The psychophysical challenge

What is needed to meet the psychophysical challenge is: (1) a paradigm for assessing perceptual interactions between the multiple stimulus dimensions; and (2) a way to model the responses obtained from observers with that paradigm which will show whether variation along one dimension influences perception of differences along another, and if so, quantifying that influence.

The Garner paradigm (Garner, 1974) was devised as a test of perceptual interaction, or *integrality*, used here to refer to the degree to which an observer's perception of one dimension of a multidimensional stimulus is influenced by the value of another, physically orthogonal dimension: Integrality contrasts with *separability*, in which the percept of one dimension does not influence the percept of the other. To apply the paradigm, a stimulus array is constructed by varying two stimulus dimensions orthogonally and listeners are required to classify various subarrays from the array.

Our previous work used a detection-theoretic model of performance in the Garner paradigm tasks (Kingston and Macmillan, 1995; Macmillan and Kingston, 1995; also Kingston and Diehl, 1995; Diehl *et al.*, submitted) to investigate perceptual integration of the acoustic correlates of pairs of potentially independent articulations. That model is also used to interpret the two experiments reported here, and to test our principal hypothesis: that the acoustic correlates of voice quality and tongue root position integrate perceptually into the flatness property.

With respect to voice quality, this hypothesis was tested quite generally, in that the stimulus arrays used in the two experiments spanned a range from very tense to very lax voice qualities. On the other hand, the range of F_1 frequencies used to simulate differences in tongue root position was confined to a rather narrow range of intermediate values. Nonetheless, this is the part of the range that is of interest, for if any vowels contrast in a language for tongue root position, it is the mid vowels (Hall *et al.*, 1974), and these vowels have F_1 values in the intermediate range used here. One might even expect that variation in voice quality could alter the flatness percept only when the vowel's F_1 is neither especially high nor low, because the more extreme F_1 values saturate the effect.

The interval between adjacent stimuli along each dimension was kept small, to a just noticeable difference, so that variation in accuracy could be used to assess the perceptual interaction between voice quality and tongue root position. Detection theory was used to construct perceptual representations of the stimulus array, and to quantify the extent to which varying one dimension distorts the percept of the other.

I. METHODS

A. Stimuli

The two experiments used similar ranges of F_1 values, but differed in what part of the voice quality continuum was paired with F_1: in the TENSE experiment, voice quality ranged from very tense to the middle of the tense–lax continuum, whereas in the LAX experiment, voice quality ranged from the middle of the tense–lax continuum to very lax. Overlap between the lax stimuli in the TENSE experiment and the tense stimuli in the LAX experiment allowed us to combine the two experiments' results in mapping the perceptual interaction between F_1 and laxness over virtually the entire tense–lax continuum.

All stimuli were synthesized with the KLSYN88 terminal analogue synthesizer (as implemented in Sensyn; Klatt and Klatt, 1990; see also Maddieson and Ladefoged, 1985; and Huffman, 1987 for data on the phonetics of voice quality contrasts). Variation in tongue root position was implemented through manipulation of F_1, and variation in voice quality (VQ) through manipulation of the percentage or quotient of the glottal cycle in which the glottis was open (OQ) and the overall tilt of the source spectrum (ST). ST is expressed in terms of the number of additional dB the source spectrum's intensity has fallen off by 3 kHz, beyond the default fall-off of -6 dB/octave. The two parts of Table I below list the four values employed for each parameter in the two experiments (see also Fig. 1). Advancement of the tongue root varies inversely (right-to-left in this table) with F_1 and is coded by the letters A–D. Laxness of the voice varies directly with OQ and ST (bottom-to-top) and is coded by the numbers 1–8.[5]

The steps along each dimension were approximately a just-noticeable difference (jnd) at 70%–80% correct, and were determined by extensive pretesting of the stimuli with the listeners from whom the data were collected. Note that a smaller range of F_1 values, with smaller steps between adjacent values was used in the TENSE than the LAX experiment. The other synthesis parameters (listed in the Appendix) were set so as to create a syllable of the shape [bVb], whose vowel was mid to high back in quality, i.e., ranging from [ɔ] as in ''baulk'' or [ʌ] as in ''buck'' to [ʊ] as in ''book.'' The vowel's steady state varied across the values in Table I.

To undermine our listeners' ability to memorize particular stimulus tokens and thus prevent them from responding to dimensions other than those varied systematically, three different variants of each stimulus were created in which the peak amplitude in the vowel's steady-state relative to the rest of the stimulus was 59, 60, or 61 dB. Each variant occurred equally often but randomly in a block of trials.

TABLE I. The 4×4 stimulus arrays used in the TENSE (bottom) and LAX (top) experiments, with the manipulated parameter values in the vowel steady states: *A* to *D*=F_1 values, and 1 to 4 and 5 to 8=Ten(se) to Int(ermediate) and Int to Lax *VQ* values for the TENSE and LAX experiments, respectively.

	LAX *VQ*		F_1			
			Advanced			Retracted
	Open quotient %	Spectral tilt dB	470	484	499	514
Lax	90	−23	*A*8	*B*8	*C*8	*D*8
	72	−17	*A*7	*B*7	*C*7	*D*7
	54	−11	*A*6	*B*6	*C*6	*D*6
Int	42	−7	*A*5	*B*5	*C*5	*D*5
	TENSE *VQ*		F_1			
			Advanced			Retracted
	Open quotient	Spectral tilt	450	477	506	536
Int	53	−11	*A*4	*B*4	*C*4	*D*4
	39	−6	*A*3	*B*3	*C*3	*D*3
	33	−4	*A*2	*B*2	*C*2	*D*2
Ten	29	−3	*A*1	*B*1	*C*1	*D*1

B. Listeners

Two different groups of eight listeners were recruited from the undergraduate student population at the University of Massachusetts, Amherst. In each experiment, listeners were chosen from a larger pool of 10–12 listeners tested. To participate in the experiment a listener was required, within 5–6 h of pretesting, to reach 70%–80% correct in classifying stimuli that differed by a single step along a single dimension for both the F_1 and *VQ* dimensions. No listener in the TENSE experiment participated in the LAX experiment. None reported any speech or hearing pathology, and all were native speakers of English. All were paid for their time. As all listeners went through at least two 2-h pretesting sessions, and were trained on each task (see Sec. II D below for task descriptions) before any data were collected, they can be considered to be well practiced with the stimuli and procedures.

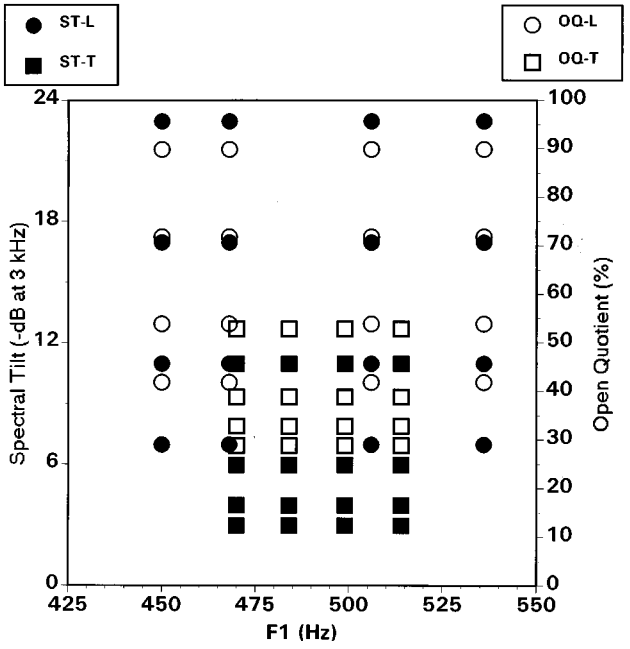

FIG. 1. Values of F_1 in Hz (horizontal axis) against spectral tilt in number of dB down at 3 kHz (left axis, filled symbols) and against open quotient in percent (right axis, open symbols) for the TENSE experiment (squares) and the LAX experiment (circles).

C. Procedures

In all tasks in both experiments, a single stimulus was presented on each trial, and the listener gave one of two responses. The listener also made a confidence judgment, on a 1–4 scale, where 1 indicated guessing, 4 certainty, and 2 and 3 intermediate levels of confidence. In all tasks, feedback as to the correct response was given at the end of the trial. A trial included the stimulus, a 2000-ms interval to give the response, a rapid triplet of tones to prompt the confidence judgment, 1500 ms to make the confidence judgment, a 500-ms feedback light, and then 2000 ms before the next trial began.

Listeners sat in a quiet room in partial isolation from other listeners in their group of four. They heard the stimuli binaurally through TDH-49 headphones, and were allowed to adjust the stimulus level individually to a comfortable and clearly audible value. They responded and gave confidence judgments by pushing appropriate buttons on a four-button response box. In addition to their responses and confidence judgments, reaction times to the responses were also recorded, to 1-ms accuracy.[6] Listeners were instructed to respond as accurately as possible, but were also told that puzzling over their answer would not help so they should also respond quickly.

In the LAX experiment, which was run first, a block of trials consisted of 16 orienting trials in which the stimuli

alternated systematically between the values relevant to the task tested in that block, followed by 96 test trials in which stimuli were presented in random order. All stimuli occurred equally often. One such block was run for tasks in which two stimulus types had to be classified, and two consecutive blocks were run for those in which four had to be. As the first six test trials were treated as further practice and were omitted from scoring, performance is assessed on the basis of 45 trials/stimulus/listener for each task in that experiment. The TENSE experiment used shorter blocks: 12 alternating orienting trails followed by 66 randomly ordered test trials. However, each condition was run twice, once early in the series of days allotted to that task type and once late. Again, two-stimulus blocks were run just once, and four-stimulus tasks twice. As the first 6 test trials in every block were again omitted from scoring, performance is assessed in the TENSE experiment from 60 trials/stimulus/listener in each task, an increase of one-third over the LAX experiment.

Each block lasted 9–12 min. Listeners heard 7–9 such blocks within each of three 1.5–2 h sessions per week. Each experiment required some 6–8 weeks to complete after initial jnd determination and training were complete.

D. Classification tasks

In both experiments, listeners classified stimuli from all 2×2 subarrays of adjacent stimuli in the 4×4 arrays; an example subarray from the TENSE experiment is stimuli *A*3, *B*3, *A*2, *B*2, and an example from the LAX experiment is stimuli C8, D8, C7, D7. In a block of trials, stimuli were classified according to one of the three rules or ''tasks'' ordinarily run in the Garner paradigm. In two of the tasks, just two stimuli were presented in a block of trials: (1) in *single-dimension* fixed classification tasks, stimuli were classified according to differences along just a single dimension, e.g., *A*3 vs *A*2 or *C*8 vs *D*8; and (2) in *correlated* fixed classification tasks, stimuli were classified according to correlated differences along both dimensions, e.g., *A*2 vs *B*3 or *C*8 vs *D*7. In the third task, *roving* classification, all four stimuli were presented in a block and were categorized according to their differences along one dimension, ignoring the other, e.g. *C*8 and *D*8 vs *C*7 and *D*7, or *A*2 and *A*3 vs *B*2 and *B*3.[7]

There are 12 single-dimension tasks along each dimension of the 4×4 array used in each experiment, e.g., for *VQ* differences *B*1 vs *B*2 (TENSE experiment) or *B*7 vs *B*8 (LAX experiment) and for F_1 differences *A*4 vs *B*4 (TENSE) or *C*6 vs *D*6 (LAX). In the correlated tasks, F_1 and tenseness of *VQ* could covary either ''negatively,'' e.g., *C*1 vs *D*2 (TENSE) or *A*7 vs *B*8 (LAX), or ''positively,'' e.g., *A*4 vs *B*3 (TENSE) or *C*6 vs *D*5 (LAX); there are nine such tasks for each correlation polarity in each 4×4 array. There are also nine roving classification tasks for each dimension in each 4×4 array, e.g., for *VQ* in the TENSE experiment, *A*2 and *B*2 vs *A*3 and *B*3 or for F_1 in the LAX experiment, *B*5 and *B*6 vs *C*5 and *C*6.

Both experiments began with the fixed classification tasks; single-dimension and correlated variants were run on alternate days. The roving classification tasks were run after all the fixed classification tasks were complete. At the beginning of each of these two phases in each experiment, listeners were trained for a day or two on the characteristics of the new tasks before any data were collected. To pseudorandomize task order, the eight listeners in each experiment were divided into two groups of four each and the tasks were run in one order for one group and in the opposite order for the other. In addition, the order in which 2×2 subarrays were run within a day was systematically mixed.

II. PSYCHOPHYSICAL ANALYSIS

A. Parallelogram models of mean integrality

1. Two-stimulus, fixed classification

As in our previous work (Kingston and Macmillan, 1995), detection theory (Green and Swets, 1966; Macmillan and Creelman, 1991) was used to model differences in our listeners' performance across tasks. The stimuli were designed to be imperfectly discriminable (70%–80% correct in single-dimension classification), and the principal measure was how accurately listeners sorted the stimuli into the classes defined by each task.

Detection theory assumes that the observer's accuracy can be represented as a map of the stimulus space onto a *decision space*, which in this application is taken to have two dimensions. Each stimulus has an average location in the space, but various sources of noise produce trial-to-trial variability on both dimensions, and the stimuli's perceptual values thus form bivariate distributions of response likelihood. In performing a specific task, the observer divides the space into regions corresponding to each response.

Because d' is a measure of the perceptual distance between the means of the corresponding distributions, in units of their standard deviation,[8] the d' values from the fixed classification tasks for a particular 2×2 subarray can be used to construct a quadrilateral representing the perceptual map of those four stimuli onto the decision space. The lengths of the sides of the quadrilateral are equal to the d' values for the four single-dimension classifications, the lengths of the diagonals to the d' values for the two correlated classifications. When the differences in performance on parallel single-dimension classifications are small (see Sec. III B below for assessment), the corresponding d' values can be averaged, so that the resulting quadrilateral is a parallelogram, as sketched in Fig. 2.

Figure 2 shows two possible maps from the stimulus to decision space. In these maps, points represent the means of the distributions, and circles contours of equal likelihood (a circle is the correct shape, for any likelihood, if the distributions are equal variance, uncorrelated, bivariate normal). The top panel's rectangular arrangement of the means in the decision space maps the case in which the perceptual value of a stimulus on one dimension does not depend on its value on the other, and the dimensions are *separable*. In the lower panel, the means of the distributions are no longer arranged rectangularly, and the value of a stimulus on one dimension does depend on its value on the other. Maddox (1992) labeled this particular violation of independence ''mean-shift integrality''; we have adopted the shorter term ''mean integrality'' (Kingston and Macmillan, 1995; Macmillan and

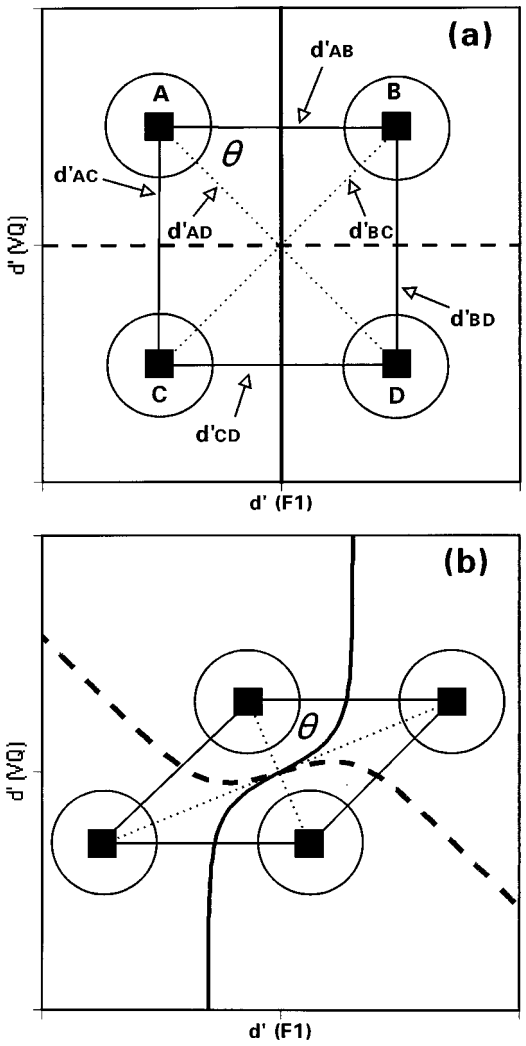

FIG. 2. (a) Rectangular versus (b) nonrectangular parallelograms for separable and mean-integral dimensions, respectively. θ represents the upper lefthand angle, which is used to estimate the degree of mean integrality. Thin, solid lines making up the sides of the parallelograms represent single-dimension d' values, thin, dotted lines the correlated d' values, as indicated in (a). Optimal criteria for the roving classification task are represented by thick lines: solid for F_1 and dashed for VQ.

Kingston, 1995). Comparing the two panels of this figure reveals that it is the (in)equality of the diagonals of the parallelogram, i.e., (in)equality of performance on the two correlated tasks in a 2×2 subarray, that determines whether the stimulus dimensions are perceptually separable or mean integral. The parallelogram representation permits calculation of a statistic that reflects the degree of mean integrality: the size of the angle of one of the parallelogram's corners; the angle in the upper left-hand corner was arbitrarily chosen. This value, referred to as θ, is 90° when the stimulus dimensions are perceptually separable but deviates from this value when they are integral. It is less than 90° when the negatively correlated task is easier than the positively correlated and greater than 90° when the difference in correlated difficulty is reversed.

According to Garner's (1974) original conception, dimensions are *separable* when classification of stimuli in which the two dimensions are correlated is no easier than single-dimension classification, and classification of stimuli with respect to one dimension while the other is roved irrelevantly is no harder than single dimension. Dimensions are integral when correlated classification is better, a *redundancy gain*, and roving classification worse, a *filtering loss*, than single-dimension classification.

In using the difference between performance on the two correlated tasks to indicate integrality, our model differs essentially from Garner's. Our models (Kingston and Macmillan, 1995; Macmillan and Kingston, 1995), following the analysis of Ashby and Townsend (1986), predict that average correlated performance will exceed average single-dimension even for the rectangular arrangement of the stimuli in the decision space, i.e., when the dimensions are separable. Simple geometric calculations show that mean correlated performance will always exceed mean single-dimension performance, but that the amount of the excess depends on both θ and the ratio of sides of the parallelogram sides.[9]

2. Filtering loss in roving classification predicted from mean integrality

The parallelogram representation can be used to predict performance on roving classification. In making this prediction, an optimal criterion is assumed, one composed of the locus of points with a likelihood ratio of 1, representing the case in which observations are equally likely to arise from either pair of distributions. (Optimal criteria are represented by thick lines in Fig. 2.) Under this assumption, no filtering loss is predicted with separable dimensions, because the optimal criterion is the same for roving classification as for single-dimension classification: a vertical or horizontal line perpendicular to the relevant dimension and parallel to the irrelevant dimension of the perceptual rectangle [Fig. 2(a)].

However, when the parallelogram is not a rectangle and the dimensions are instead mean integral [Fig. 2(b)], the optimal criteria in roving classification are no longer perpendicular to the stimulus dimensions. Instead, the optimal criterion for, say, roving classification by F_1 differences [thick solid line in Fig. 2(b)] approaches the perpendicular bisector of the line between the means of response distributions at the top of the parallelogram for large positive values of VQ, passes through the center of gravity of the parallelogram, and approaches the perpendicular bisector of the line between the means of the response distributions at the bottom of the parallelogram for large negative values of VQ. The criterion for roving classification by VQ differences [thick dashed line in Fig. 2(b)] similarly passes through the parallelogram's center of gravity and approaches the bisectors of the right and left sides of the parallelogram for large positive and negative values of F_1.

According to these optimal rules, roving classification should not be as good as in the corresponding single-dimension conditions. The size of the decline in performance resulting from mean integrality can be predicted for any parallelogram (Macmillan and Kingston, 1995): the magnitude of the predicted decline increases as θ diverges from 90°, and the effect is greatest when the two single-dimension sensitivities are equal.

TABLE II. Mean d' (se) across listeners for single-dimension *VQ* classification. In this and all subsequent tables, a space separates the results of the LAX experiment at the top from those of the TENSE experiment at the bottom.

	Low	F_1 level		High
VQ pair	*A*	*B*	*C*	*D*
7–8	1.72 (0.33)	1.30 (0.24)	1.26 (0.25)	0.96 (0.31)
6–7	3.10 (0.27)	2.73 (0.21)	2.69 (0.30)	3.14 (0.49)
5–6	1.52 (0.39)	0.98 (0.34)	0.53 (0.14)	0.14 (0.07)
3–4	3.07 (0.25)	3.21 (0.28)	2.63 (0.28)	2.13 (0.31)
2–3	1.18 (0.26)	1.98 (0.31)	1.98 (0.25)	1.05 (0.17)
1–2	1.15 (0.19)	1.62 (0.20)	1.50 (0.19)	1.44 (0.19)

Garner (1974) also predicts a decline in performance in roving classification as compared to fixed classification for integral dimensions, but for different reasons. In his model, irrelevant variation of one of two integral dimensions *interferes* with the observer's ability to detect that stimuli actually have the same value with respect to the relevant dimension. In other words, integrality interferes with ''filtering'' out of the perceptual effects of a stimulus value for the irrelevant dimension in judging its value for the relevant one. Moreover, Garner's model predicts filtering losses will be equally large for all the stimuli in the array, whereas the parallelogram model predicts an asymmetry: with the perceptual map in Fig. 2(b), filtering loss will be greater for stimuli *A* and *D* than for *B* and *C*. The extent of this asymmetry increases with the degree of mean integrality.

3. Testability of accuracy models

Assuming that the same perceptual representation underlies decisions in all tasks run with a particular 2×2 subarray and that decision criteria are optimal, performance in one correlated and both roving classifications can be predicted from performance in the single-dimension tasks and the other correlated condition. In addition, the more mean integral the dimensions are, the greater filtering loss for stimuli along the shorter than the longer diagonal in roving classification, so measuring this asymmetry serves as a further test of the model. The method introduced by Marascuilo (1970) for determining whether two d' values differ significantly from one another is also used throughout the following discussion in assessing the model's representation of these data, at the level of individual listeners.

B. Response times and their relation to perceptual distance

In most applications of the Garner paradigm, differences in task performance and thus the separable-integral question are assessed in terms of differences in response time (RT) rather than accuracy because the stimuli in single-dimension classification differ by many jnds rather than just one. Although the use of RTs leads naturally to interpretation of the results in terms of processes rather than the representations, Ashby, Maddox, and their colleagues (Ashby and Maddox, 1994; Ashby *et al.*, 1994) have recently put forward the ''RT-distance hypothesis''; time to respond in classifying a single stimulus is an inverse function of its distance from the criterion. The RT-distance hypothesis predicts that RTs and d' values correlate inversely, a prediction tested below.

III. RESULTS

A. Performance as a function of dimensions, tasks and subarrays

1. Accuracy

Tables II and III show mean performance on single-dimension classification for *VQ* and F_1, respectively, for each 2×2 subarray in the two experiments. The differences in performance in classifying individual stimulus pairs for differences on these dimensions are to some extent a function of their physical separation in the stimulus parameter space: differences in the voice quality parameters account for 22.4% of the variance in performance on *VQ* classification and differences in F_1 for 37.5% of the variance on F_1 classification. Preliminary stimulus evaluation clearly overestimated jnds for some stimulus pairs. An advantage of our detection-theoretic approach (Sec. III B) is that exactly equal perceptual spacing is unnecessary.

Table IV shows the results of the roving classification task for both *VQ* and F_1 differences. Performance varies between 2×2 subarrays in essentially the same way as in single-dimension classification, but comparison with Tables II and III shows a noticeable decline: for *VQ* differences roving performance is 0.67 of single-dimension in the TENSE experiment and 0.53 in the LAX experiment; for F_1 differences these proportions are 0.61 and 0.77, respectively.

Performance on the positively and negatively correlated tasks is displayed separately in each pair of cells in Table V.

TABLE III. Mean d' (se) across listeners for single-dimension F_1 classification.

		F_1 pair		
VQ level		*A*–*B*	*B*–*C*	*C*–*D*
Lax	8	2.23 (0.38)	1.70 (0.19)	2.88 (0.43)
	7	3.29 (0.48)	2.74 (0.37)	2.61 (0.39)
	6	1.76 (0.37)	2.84 (0.48)	2.52 (0.42)
	5	2.94 (0.60)	2.31 (0.41)	2.87 (0.40)
	4	2.02 (0.28)	1.05 (0.23)	0.93 (0.19)
	3	2.08 (0.20)	2.52 (0.21)	1.77 (0.28)
	2	2.69 (0.25)	1.97 (0.22)	1.21 (0.25)
Tense	1	2.85 (0.25)	1.67 (0.26)	1.14 (0.22)

TABLE IV. Mean d' (se) across listeners for VQ and F_1 roving classification.

	F_1 pair					
	$A-B$		$B-C$		$C-D$	
VQ pair	VQ	F_1	VQ	F_1	VQ	F_1
7–8	0.94 (0.38)	3.25 (0.46)	1.05 (0.21)	2.17 (0.35)	0.70 (0.12)	2.18 (0.39)
6–7	1.47 (0.20)	1.56 (0.23)	1.08 (0.25)	1.19 (0.36)	1.57 (0.11)	1.57 (0.33)
5–6	0.50 (0.21)	1.17 (0.25)	0.47 (0.18)	1.87 (0.32)	0.18 (0.11)	2.90 (0.48)
3–4	2.47 (0.29)	0.90 (0.21)	2.17 (0.28)	0.44 (0.11)	1.77 (0.32)	0.52 (0.09)
2–3	0.68 (0.16)	1.60 (0.21)	1.06 (0.19)	1.59 (0.17)	0.79 (0.11)	0.85 (0.15)
1–2	0.58 (0.12)	1.77 (0.19)	0.83 (0.12)	1.30 (0.26)	1.13 (0.15)	0.95 (0.17)

Performance is noticeably better when tenseness of voice quality correlates positively with F_1 than when these dimensions correlate negatively, except for intermediate voice qualities (row pairs 3–4 and 5–6).

2. Correlation between accuracy and response times

The correlation between mean RTs and d' values across all tasks and subarrays is -0.79, indicating a very strong tendency for more accurate responses to be faster. A straight line fitted to a plot of mean RT values against mean d' values indicates that RT is reduced 85 ms for each unit increase in d'; this line accounts for 62% of the variance in RT values. These results thus accord well with Ashby and Maddox's RT-distance hypothesis.

B. Parallelogram models of fixed classification

Parallelogram representations of the data for each 2×2 subarray are displayed in Fig. 3. In each panel, the vertical sides of the parallelogram have length d'_{VQ} (the average d' in the two VQ fixed classifications for the same subarray) and the other two sides have length d'_{F_1} (the average d' in the two F_1 fixed classifications). The correlated d' values correspond to the diagonals of the parallelogram. Iteration was used to find the value of θ that provided the best fit to all six d' values.

The extent to which the correlated and single-dimension data fit the same parallelogram is one test of the model. Each observed and predicted d' was converted to proportion correct $p(c)$ (assuming no bias) and the rms error computed; rms errors are listed next to the value of θ in each panel of Fig. 3. By this measure, the fit of the parallelogram model was good: in $p(c)$ units, rms errors averaged just 0.020 in the LAX experiment and 0.011 in the TENSE experiment. The halving of the error in the latter probably resulted from the one-third increase in the number of trials per point and the more extensive pretesting and training of listeners.

The appropriateness of fitting a parallelogram (rather than an irregular quadrilateral) can be assessed, by comparing d' values for parallel single-dimension tasks, because fitting a parallelogram assumes they should be equally easy. As there are nine 2×2 subarrays to which parallelograms could be fit and eight listeners in each experiment, each allows 72 comparisons for each dimension, VQ and F_1. Assuming $\alpha=0.05$, in the TENSE experiment, d' values for parallel VQ tasks were significantly different four (6%) times, and for F_1 eight (11%) times; in the LAX experiment, there were eight (11%) significant differences for both VQ and F_1. (If $\alpha=0.10$, these scores are 7, 11, 9, and 17, respectively.) Therefore, in just under 10% of the cases overall, the parallelogram assumption is incorrect (if $\alpha=0.10$, this rises to just over 15%). Violations do not cluster in any region in either $F_1\times VQ$ array. This success suggests that the parallelogram models provide a good fit to the data.

The observed values of θ, the degree of mean integrality, are either less than or greater than 90°, indicating mean integrality of F_1 and VQ. In the TENSE experiment, 24/72 (33%) comparisons of positively versus negatively correlated d' values were significantly different at $\alpha=0.05$, an additional 5 more at $\alpha=0.10$ [total 29 (40%)], and in the LAX experiment, 31 (43%) were significantly different at $\alpha=0.05$, an additional 2 more at $\alpha=0.10$ [total 33 (46%)].

As would be expected from the figure (but from not any *a priori* theorizing), the dimensions integrate quite differently at more tense or lax than at intermediate voice quali-

TABLE V. Mean d' (se) across listeners for positively and negatively correlated classification.

	F_1 pair					
	$A-B$		$B-C$		$C-D$	
VQ pair	Positive	Negative	Positive	Negative	Positive	Negative
7–8	3.80 (0.34)	2.15 (0.36)	4.10 (0.26)	3.57 (0.22)	3.71 (0.37)	1.45 (0.40)
6–7	3.80 (0.33)	0.57 (0.27)	4.51 (0.03)	0.61 (0.32)	4.27 (0.16)	1.06 (0.40)
5–6	1.45 (0.40)	3.86 (0.28)	3.28 (0.40)	4.28 (0.09)	2.72 (0.54)	3.73 (0.27)
3–4	2.73 (0.29)	3.62 (0.24)	2.29 (0.23)	3.71 (0.22)	1.79 (0.24)	3.15 (0.31)
2–3	3.32 (0.22)	2.99 (0.25)	3.45 (0.20)	2.07 (0.24)	2.87 (0.27)	1.13 (0.21)
1–2	3.98 (0.09)	2.20 (1.31)	3.43 (0.20)	1.09 (0.15)	2.50 (0.23)	1.27 (0.20)

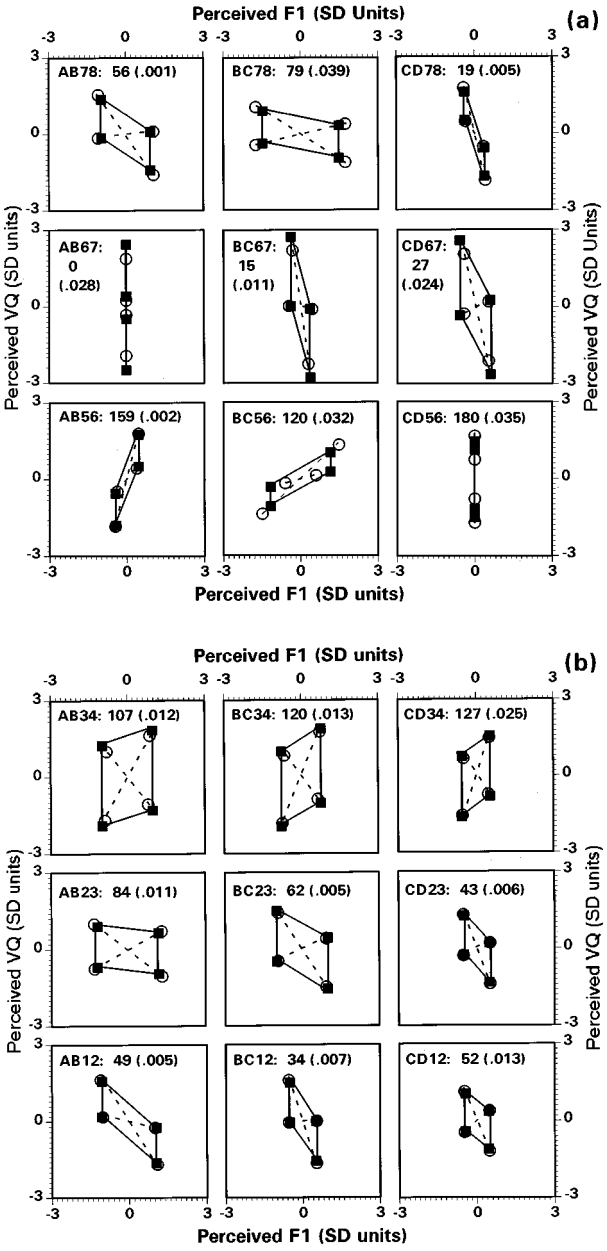

FIG. 3. Parallelograms fitted to single-dimension (filled squares) and correlated (open circles) classification data for each 2×2 subarray in the 4×4 arrays of the (a) LAX and (b) TENSE experiment. The letter–number pairs identifying each subarray denote their F_1 and VQ values, respectively. Solid lines represent the sides and dashed lines the diagonals of the parallelograms. The first number listed after the identifier is the θ (in degrees) for the parallelogram that best fits the two estimates. The rms errors in parentheses represent the discrepancy, in units of proportion (correct), between the baseline and correlated estimates of the locations of the corners of the parallelogram.

ties. The value of θ is systematically less than 90° for the more extreme voice qualities but greater than 90° for the intermediate ones: row pairs 1–2, 2–3 and 6–7, 7–8 [Fig. 3(a), AB67-CD78, Fig. 3(b) AB12–CD23] versus row pairs 3–4 and 5–6 [Fig. 3(a) AB56-CD56, Fig. 3(b), AB34–CD34]. Significant differences in positively versus negatively correlated d' values are also rarer for the intermediate than extreme voice qualities: in the TENSE experiment, row pairs 1–2 and 2–3 yielded 13/27 (48%) and 7 (26%) significant differences (α=0.05), respectively, whereas row pair 3–4 yielded only 3 (11%), and in the LAX experiment, row pair 5–6 yielded only 6 (22%) significant differences, whereas row pairs 6–7 and 7–8 yielded 19 (70%) and 6 (22%), respectively. VQ and F_1 are apparently more separable at intermediate than extreme voice qualities, although they are more nearly separable at the laxest voice qualities, too.

That θ>90° for the tenser and laxer voice qualities indicates that stimuli in which tenseness and F_1 covary positively are easier to classify at the two ends of the voice quality continuum. However, for intermediate voice qualities the direction of mean integrality is reversed with respect to either tenser or laxer voice qualities, because *negative* covariation between tenseness and F_1 makes classification easier in this part of the continuum. These reversals compared to more extreme voice qualities are far stronger in the LAX than TENSE experiment, despite the overlap in voice quality parameters values for the intermediate voice qualities in the two 4×4 arrays (see Table I and Fig. 1). In summary, our listeners apparently integrate F_1 and VQ differences in the predicted way, i.e., into the perceptual property called ''flatness'' at the more tense and lax extremes, but not in the middle of the tense–lax continuum.

C. Using the parallelograms to predict roving classification performance

1. Predicted versus observed filtering loss

Mean integrality predicts that performance on roving classification will be worse than that on single-dimension classification. When d' values obtained in roving classification are compared to the mean d' values obtained in the analogous single-dimension classifications, in the TENSE experiment 13/72 (18%) cases of VQ roving classification showed a significant (α=0.05) filtering loss as did 19 (26%) cases of F_1 classification; in the LAX experiment, significant filtering losses were obtained in 15 (21%) and 18 (25%) cases of VQ and F_1 roving classification, respectively. Overall, significant filtering losses were obtained in 23% of the cases. (At α=0.10, 18, 21, 19, and 20 cases, respectively, of significant filtering were obtained, a total of 27% overall.)

The loss predicted from the degree of mean integrality [in proportion correct $p(c)$ units] is listed in Table VI, together with the discrepancy between the observed and predicted loss. Predicted losses and discrepancies are listed for each dimension within each 2×2 subarray separately.

The magnitude of the predicted loss tends to be small, averaging in the TENSE experiment just 0.017 for F_1 and 0.018 for VQ and in the LAX experiment just 0.047 for F_1 and 0.060 for VQ. The amount of loss beyond that predicted from mean integrality is by contrast relatively large, in the TENSE experiment averaging 0.111 for F_1 and 0.081 for VQ and in the LAX experiment 0.048 for F_1 and 0.050 for VQ. Thus, the parallelogram model accounts for about 15% of the filtering loss in the TENSE experiment and about 52% in the LAX experiment. The additional loss may arise because roving variation along the irrelevant dimension increases the stimulus's variability and thus the listener's uncertainty about its value for the relevant dimension. This *variance integrality* is very like the interference to which

TABLE VI. Predicted loss in roving classification from mean integrality, and the discrepancy (in parentheses) between predicted and observed loss [both in $p(c)$ units]. A negative predicted loss indicates that roving is expected to be easier than single-dimension classification. A positive discrepancy indicates a loss greater than predicted, a negative one a lesser loss than predicted. Each 2×2 subarray is represented by a pair of cells, one above the other, representing predicted loss and discrepancies for the two dimensions of classification, cf. Table IV.

Exp.	VQ pair	Relevant dimension	F_1 pair A:B	B:C	C:D
LAX	7:8	F_1	0.017 (−0.050)	0.005 (0.036)	0.033 (0.017)
		VQ	0.021 (0.071)	−0.001 (0.070)	0.083 (−0.007)
	6:7	F_1	0.121 (0.001)	0.117 (0.070)	0.086 (0.035)
		VQ	0.112 (0.047)	0.124 (0.075)	0.084 (0.059)
	5:6	F_1	0.035 (0.157)	0.003 (0.114)	0.007 (0.001)
		VQ	0.084 (0.048)	0.010 (0.068)	0.026 (0.022)
TENSE	3:4	F_1	0.006 (0.169)	0.021 (0.220)	0.022 (0.172)
		VQ	0.002 (0.031)	0.012 (0.035)	0.015 (0.053)
	2:3	F_1	−0.007 (0.114)	0.010 (0.067)	0.024 (0.084)
		VQ	−0.001 (0.163)	0.018 (0.113)	0.028 (0.096)
	1:2	F_1	0.017 (0.094)	0.039 (0.044)	0.017 (0.034)
		VQ	0.028 (0.115)	0.045 (0.082)	0.015 (0.042)

Garner attributed any filtering loss (see also Maddox, 1992, on ''variance-shift integrality'') and will be explored in a subsequent paper.

2. Asymmetries in roving classification

The parallelogram model not only predicts a filtering loss in roving classification when the stimulus dimensions are mean integral but also a greater loss for the stimuli at the ends of the shorter than the longer diagonal. This asymmetry should furthermore be a function of θ's deviation from 90°.

Tables VII and VIII list d' values calculated separately for the stimuli along the two diagonals of each parallelogram for roving classification by VQ and F_1 differences, respectively. If the parallelograms in Fig. 3 correspond to the listener's perceptual representation of the stimuli in the 2×2 subarrays across tasks, then the differences in performance in roving classification between stimuli along the positive and negative diagonals should be in the same direction as the differences in the corresponding correlated tasks (Table V). This prediction holds for both dimensions because stimuli along the diagonals must be assigned to different classes in both roving classifications. The prediction is strongly upheld. In roving classification for F_1 differences (Table VIII), the asymmetries correspond to those in the correlated tasks in every 2×2 subarray in both experiments. The correspondence is nearly as good in roving classification for VQ differences (Table VII), where only 1 out of 18 2×2s shows the opposite asymmetry in roving than correlated classification: $C-D$×5–6 in the LAX experiment. In addition, the size of the asymmetry between positive and negative performance in roving classification correlates well with the deviation of θ from 90°: the correlation is 0.81 for F_1 and VQ differences combined, 0.87 for F_1 alone, and 0.77 for VQ alone. These results strongly support the contention that listeners employ the same perceptual representation in roving as fixed classification.

IV. DISCUSSION

The results of these experiments mix both a psychophysical and a phonetic loss among many wins.

Among the psychophysical wins can be counted the evidence that listeners use the same perceptual representations in roving as fixed classification, the very successful fit of the parallelogram models, the close agreement between RTs and accuracy, which qualitatively supports Ashby and Maddox's (1994) RT-distance hypothesis, and similarity in the listen-

TABLE VII. Mean d' (se) across listeners for positively and negatively correlated pairs in roving classification for VQ differences.

VQ pair	F_1 pair $A-B$ Positive	$A-B$ Negative	$B-C$ Positive	$B-C$ Negative	$C-D$ Positive	$C-D$ Negative
7–8	1.23 (0.37)	1.00 (0.45)	2.16 (0.45)	0.48 (0.25)	2.16 (0.28)	−0.07 (0.27)
6–7	3.15 (0.30)	0.41 (0.20)	2.51 (0.38)	0.27 (0.24)	3.01 (0.46)	0.49 (0.15)
5–6	0.10 (0.17)	1.01 (0.48)	0.64 (0.19)	0.77 (0.31)	0.65 (0.26)	−0.15 (0.25)
3–4	2.24 (0.29)	3.01 (0.28)	1.95 (0.36)	2.78 (0.29)	1.64 (0.30)	2.10 (0.38)
2–3	1.36 (0.22)	0.28 (0.20)	1.88 (0.26)	0.86 (0.24)	1.73 (0.20)	0.39 (0.18)
1–2	1.84 (0.31)	−0.26 (0.15)	1.51 (0.27)	0.19 (0.14)	1.60 (0.24)	0.71 (0.18)

TABLE VIII. Mean d' (se) across listeners for positively and negatively correlated stimulus pairs in roving classification for F_1.

	F_1 pair					
	A–B		*B–C*		*C–D*	
VQ pair	Positive	Negative	Positive	Negative	Positive	Negative
7–8	3.30 (0.40)	3.14 (0.41)	2.79 (0.33)	1.93 (0.35)	3.13 (0.44)	1.61 (0.33)
6–7	3.42 (0.31)	0.70 (0.27)	2.90 (0.28)	0.72 (0.28)	2.56 (0.34)	0.69 (0.37)
5–6	0.49 (0.18)	2.75 (0.46)	1.39 (0.35)	2.61 (0.45)	2.67 (0.46)	3.02 (0.41)
3–4	0.52 (0.29)	1.55 (0.24)	0.24 (0.22)	1.42 (0.18)	0.29 (0.29)	1.39 (0.29)
2–3	2.46 (0.30)	1.34 (0.29)	2.55 (0.26)	0.95 (0.20)	1.70 (0.26)	0.41 (0.12)
1–2	2.91 (0.24)	1.14 (0.22)	2.06 (0.34)	0.81 (0.26)	1.95 (0.26)	0.31 (0.17)

ers' responses to the overlapping portions of the two experiments' stimulus arrays. The most serious loss was that the parallelogram model substantially underestimated the filtering loss, especially in the TENSE experiment.

The principal phonetic win is finding the predicted direction of mean integrality at the extreme voice qualities, where positive correlation between tenseness and F_1 enhanced discriminability. For two-thirds of the stimuli in both experiments, voice quality and tongue root position apparently integrate perceptually into the property flatness. However, flatness differences clearly do not predict performance at intermediate voice qualities, as shown by the marked reversals in the direction of mean integrality there compared to laxer or tenser voice qualities.

A closer examination of the acoustics of the stimuli and their relation to listeners' performance hints at the basis for these reversals but does not explain them away. A useful acoustic correlate of flatness is the intensity differences (in

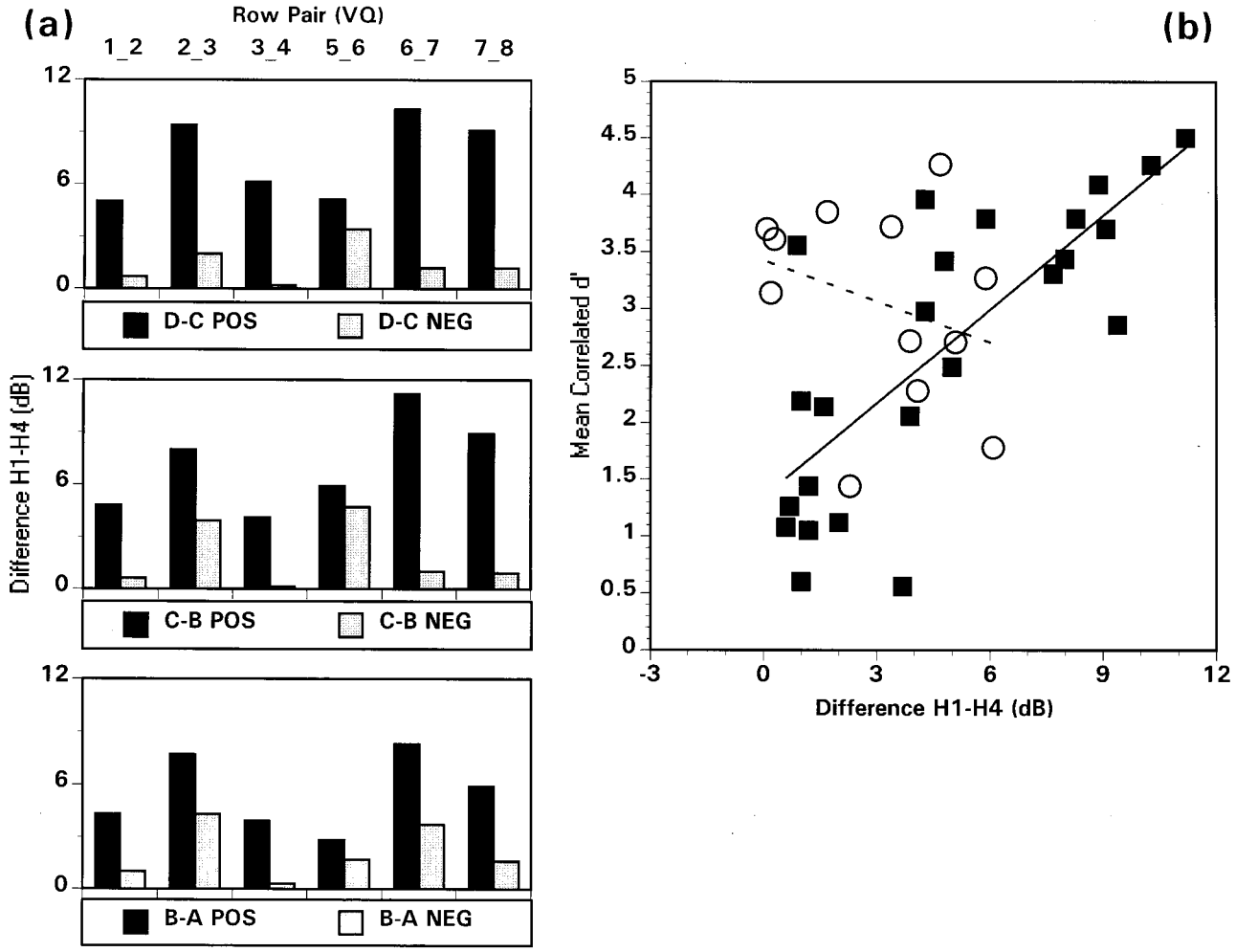

FIG. 4. (a) Difference between positively (black bars) and negatively (gray bars) correlated stimuli in the intensity difference between the first and fourth harmonics (H_1-H_4), in dB; F_1 decreases from top-to-bottom. (b) Mean d' values by differences in H_1-H_4. Stimulus pairs with extreme, i.e., tenser and laxer voice qualities (row pairs: 1–2, 2–3, 6–7, and 7–8) are plotted with filled squares, those with *intermediate* voice qualities (3–4 and 5–6) are plotted with open circles. A separate line is fitted to the *extreme* (solid) and intermediate (dashed) data in each case.

dB) between the first and fourth (H_1-H_4) harmonics, extracted from 1024 point (102.4 ms) FFTs centered on each vowel's midpoint (the center of gravity of the 0–1 kHz frequency interval in this spectrum was a similar measure). A laxer voice quality raises the intensity of H_1 relative to higher harmonics and a more advanced tongue root lowers F_1; the latter will affect H_4's intensity particularly, as that harmonic was closest in frequency to F_1. Thus, H_1-H_4 will be larger the flatter the vowel, and if flatness differences influence the listeners' performance then vowel pairs which differ more in this (or an analogous flatness measure such as center of gravity) should be easier to classify.

Figure 4(a) is a plot of magnitude of the difference in H_1-H_4 values between the members of positively versus negatively correlated vowel pairs in the two experiments, and shows that in all cases the positively correlated stimuli differ more than the negatively correlated ones in their H_1-H_4 values. This figure also shows, however, that the difference in magnitude is uniformly smaller at the intermediate (3–4 and 5–6) voice qualities than at tenser or laxer extremes. These smaller differences predict that if listeners were attending to flatness differences, performance on the negatively correlated task should be closer to that on the positively correlated task; i.e., weaker mean integrality or even separability should have been obtained at intermediate voice qualities. Fewer significant differences between positively and negatively correlated performance were found for intermediate voice than extreme voice qualities, but a problem remains.

By this acoustic measure, mean integrality should have been weaker for stimuli in the 5–6 row pair of the LAX experiment than in the overlapping 3–4 row pair in the TENSE experiment, but the results were quite the reverse. Not only was more evidence of separability obtained in the 3–4 row pair but the 5–6 row pair yielded exceptionally strong evidence of mean integrality in the *opposite* direction; i.e., θ values were markedly greater than 90°.

Figure 4(b), which plots mean d' values in the correlated tasks against the corresponding H_1-H_4 differences, shows the independence of correlated performance from flatness differences in another way. When just the mean d' values for extreme stimulus pairs (row pairs: 1–2, 2–3, 6–7, and 7–8; filled figures and solid line) are regressed against these differences, the percentage of variance accounted for is 58.9%. Mean d' increases by 0.274 units for each dB increase in the H_1-H_4 difference between the members of a stimulus pair. On the other hand, less than 1% of the variance is accounted for by the dashed line fitted to mean d' by H_1-H_4 differences for the intermediate voice qualities (row pairs: 3–4 and 5–6, open figures and dashed line).

This acoustic exploration shows that listeners probably only attend to flatness in judging these stimuli at the two ends of the voice quality continuum but not in its middle. These results furthermore predict that languages in which voice quality covaries with tongue root position would use either tense or lax extremes, because it is at the extremes that the flatness differences are greatest.[10] Because the acoustic correlates of voice quality differences interact with those of tongue root advancement so as to enhance the distinctiveness of vowels, their deliberate covariation is advantageous to speakers (and listeners).

Further work on the perceptual interaction between the acoustic correlates of voice quality and tongue root position is currently underway, to address some of the questions left unanswered in the current paper, and in hopes of turning what remain losses into wins. A subsequent paper will report how varying the range of variation of the relevant dimension, in complete identification of all four values of that dimension, affects accuracy compared to the two values that had to be identified in single-dimension fixed classification. Examining identification of the entire range of stimuli in an array may shed light on the reversal of mean integrality in the middle of the voice quality continuum. That paper will also examine the effects on filtering loss of increasing the range of irrelevant variation from just two to four values, and will thereby provide an account of variance integrality.

ACKNOWLEDGMENTS

We gratefully acknowledge the support for the work reported here obtained through Grant No. R-29-DC01708-2 from the National Institute of Deafness and Communicative Disorders, National Institutes of Health to the first author, and Grant No. DBS92-12043 from the National Science Foundation to the second author. Preliminary versions of this work were presented at the 127th and 129th meetings of the Acoustical Society of America (Thorburn *et al.*, 1994; Walsh *et al.*, 1995), at the 13th International Congress of Phonetic Sciences, Stockholm (Kingston *et al.*, 1995), to the Sound Seminar in the Linguistics Department, University of Massachusetts, and to the Linguistics Research group at AT&T Bell Laboratories; the comments of the audiences at all these presentations contributed positively to the preparation of this paper. In addition, the comments of our colleague José Benkí were very helpful at all stages in the writing of this paper. Finally, the comments of Keith Johnson and an anonymous reviewer have substantially improved this paper.

APPENDIX

Tables AI and AII below list the parameters that did not vary across the stimuli used in stimuli. Table AI shows the amplitude of voicing (AV) and F_0 profile, and Table AII the formant frequencies. A plateau occurs between points specified for the same value in adjacent cells in these tables, linear interpolation between points specified for different values in adjacent cells.

TABLE AI. Profile for amplitude of voicing (AV) and F_0 parameters for the stimuli used in both experiment I and experiment II. The middle row indicates roughly the correspondence of the inflection points in these profiles with the segments of the utterance.

Time (ms)	0	10	80	90	270	280	355	365
AV (dB)	0	48	48	60	60	48	48	0

Segments	*b*		*V*		*b*	
Time (ms)	0	80	115	180	280	365
F_0 (Hz)	120	120	140	140	120	120

TABLE AII. Profiles of open quotient (OQ) in percent of glottal cycle, spectral tilt ($-$dB @ 3 kHz), formant frequencies for F_1-F_6, and corresponding bandwidths.

Time (ms)	0	75	115	245	280	365
OQ (%)	75	75			75	75
ST ($-$dB)	39	39	see Table I		39	39
F_1 (Hz)	180	180			180	180
F_2	700	700	1070	1070	700	700
F_3	1900	1900	2300	2300	1900	1900
F_4	3100	3100	3400	3400	3100	3100
F_5			3700			
F_6			4990			

Segments	*b*		*V*		*b*	
Time (ms)	0	75	90	265	280	365
B_1 (Hz)	500	500	60	60	500	500
B_2	1000	1000	90	90	1000	1000
B_3	1000	1000	150	150	1000	1000
B_4	1000	1000	200	200	1000	1000
B_5	1500	1500	200	200	1500	1500
B_6			4000			

[1]Local (1995) reports that in Kalenjin (another Nilotic language) the pattern of covariation described above is reversed: a laxer voice quality occurs with retracted tongue root and vice versa. This counterexample does not nonetheless undermine the case made here, as a great many languages have been reported to have the pattern of covariation for which we seek an explanation. And in contrast to Local's description of Kalenjin, Tucker (1966) describes the advanced tongue root vowels as being produced with breathy voice and the retracted tongue root vowels with creaky voice in the Nandi–Kipsigis dialect of Kalenjin, and Hall *et al.* (1974) also report creaky voice as a correlate of the retracted tongue root vowels in the Elgyeo dialect. Although both reports rely on impressionistic rather than instrumental evidence, the difference between these descriptions and Local's suggests that at least some speakers of Kalenjin produce the more common covariation between tongue root position and voice quality.

[2]Another physiological linkage is the connection between the hyoid bone and the superior edges and superior cornua of the thyroid laminae. It is hard to imagine, however, how pulling the cornua forward when the hyoid bone is advanced in the tongue root would lax the voice.

[3]''Flat(ness)'' is obviously used here in a different sense than that of Jakobson *et al.* (1952) feature (flat), which represented the perceptual effect of the lowering of the second and higher formant frequencies brought about by the secondary articulations, labialization, velarization, or pharyngealization. See also Ohala (1985).

[4]We are grateful to Keith Johnson for making clear the necessity of discussing this third kind of explanation.

[5]Despite the large range of voice qualities used in each experiment, no listener reported, during the course of the experiment nor in debriefing afterwards, that any stimuli were unnatural or nonspeech-like.

[6]The clock started with the beginning of the stimulus, so RTs include the 365 ms of the stimulus. Also, any responses $\leqslant$100 ms or $\geqslant$2800 ms were excluded from the evaluation of speed or accuracy of response.

[7]Single-dimension fixed classification corresponds to the baseline task in traditional Garner nomenclature, and roving classification to the selective attention task; the correlated task is the same. The more revealing names used here are based on nomenclature introduced by Durlach *et al.* (1989).

[8]The d' value is expressed in standard deviation units because it is calculated from the z-score equivalents of the hit and false alarm proportions; see Macmillan and Creelman (1991).

[9]The 60–120° range of θ's predicts better mean correlated performance only if all four sides of the parallelogram are equal; i.e., the sides ratio as well as θ determine whether mean correlated performance will exceed single dimension performance (see Ashby and Townsend, 1986; Maddox, 1992; Macmillan and Kingston, submitted, for discussion and exemplification).

[10]Perhaps the Kalenjin speakers examined by Local (1995) use the middle of the voice quality continuum and may therefore reverse the direction of covariation observed elsewhere.

Ashby, F. G., Boynton, G., and Lee, W. W. (**1994**). ''Categorization response time with multidimensional stimuli,'' Percept. Psychophys. **55,** 11–27.

Ashby, F. G., and Maddox, W. T. (**1994**). ''A response time theory of separability and integrality in speeded classification,'' J. Math. Psychol. **38,** 423–466.

Ashby, F. G., and Townsend, J. T. (**1986**). ''Varieties of perceptual independence,'' Psychol. Rev. **93**, 154–179.

Baer, T., Alfonso, P. J., and Honda, K. (**1988**). ''Electromyography of tongue muscles during vowels in /əpVp/ environment,'' Ann. Bull. Res. Inst. Logoped. Phoniat., University of Tokyo, **7**, 7–18.

Bloedel, S. L. (**1994**). ''An analysis of the acoustic correlates of breathy phonation in the speech of adult men and women and pre-pubescent males,'' M.S. thesis, University of Wisconsin, Madison.

Denning, K. (**1989**). ''The diachronic development of phonological voice quality,'' Ph.D. dissertation, Stanford University.

Diehl, R. L., and Kingston, J. (**1991**). ''Phonetic covariation as auditory enhancement: The case of the [+voice]/[−voice] distinction,'' Perilus **15**, 139–143.

Diehl, R. L., Kingston, J., and Castleman, W. A. (**submitted**). On the internal perceptual structure of distinctive features: The [voice] distinction, J. Acoust. Soc. Am.

Durlach, N. I., Tan, H. Z., Macmillan, N. A., Rabinowitz, W. R., and Braida, L. D. (**1989**). ''Resolution in one dimension with random variations in background dimensions,'' Percept. Psychophys. **46,** 293–296.

Fowler, C. A. (**1996**). ''Listeners do hear sounds, not tongues,'' J. Acoust. Soc. Am. **99,** 1730–1741.

Garner, W. R. (**1974**). *The Processing of Information and Structure* (Erlbaum, Potomac, MD).

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hall, B. L., Hall, R. M. R., Pam, M. D., Myers, A., Antell, S. A., and Cherono, G. K. (**1974**). ''African vowel harmony systems from the vantage point of Kalenjin,'' Afrika und Übersee **57**, 241–267.

Hoemeke, K., and Diehl, R. L. (**1994**). ''Perception of vowel height: The role of $F1$-$F0$ distance,'' J. Acoust. Soc. Am. **96**, 661–674.

Huffman, F. E. (**1976**). ''The register problem in fifteen Mon–Khmer languages,'' in *Austroasiatic Studies I, Oceanic Linguistics Publication 15*, edited by P. N. Jenner, L. C. Thompson, and S. Starosta (University Press of Hawaii, Honolulu), pp. 575–590.

Huffman, M. K. (**1987**). ''Measures of phonation type in Hmong,'' J. Acoust. Soc. Am. **81,** 495–504.

Jackson, M. T. T. (**1988**). ''Phonetic Theory and Cross-Linguistic Variation in Vowel Production,'' University of California, Los Angeles, Working Papers in Phonetics No. 71.

Jacobsen, L. C. (**1978**). ''DhoLuo Vowel Harmony: A Phonetic Investigation,'' Univ. California, Los Angeles, Working Papers in Phonetics, No. 43.

Jacobsen, L. C. (**1980**). ''Voice quality harmony in Western Nilotic languages,'' in *Issues in Vowel Harmony*, edited by R. M. Vago (John Benjamins B. V., Amsterdam).

Jakobson, R., Fant, G., and Halle, M. (**1952**). *Preliminaries to Speech Analysis* (MIT, Cambridge, MA).

Kingston, J. (**1991**). ''Integrating articulations in the perception of vowel height,'' Phonetica **48**, 149–179.

Kingston, J., and Diehl, R. L. (**1994**). ''Phonetic knowledge,'' Language **70**, 419–454.

Kingston, J., and Diehl, R. L. (**1995**). ''Intermediate properties in the perception of distinctive feature values,'' in *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence*, edited by B. Connell and A. Arvaniti (Cambridge U.P., Cambridge, U.K.), pp. 7–27.

Kingston, J., and Macmillan, N. A. (**1995**). ''Integrality of nasalization and F_1 in vowels in isolation and before oral and nasal consonants: A detection-theoretic application of the Garner paradigm,'' J. Acoust. Soc. Am. **97,** 1261–1285.

Kingston, J., Walsh, L. J., Bartels, C. Thorburn, R., and Macmillan, N. A. (**1995**). ''Integrating voice quality and tongue root position in perceiving vowels,'' in *Proceedings of the 13th International Congress of Phonetic Sciences*, edited by K. Elenius and P. Branderud, Vol. 2 (Stockholm), pp. 514–517.

Klatt, D. H., and Klatt, L. (**1990**). ''Analysis, synthesis, and perception of voice quality variations among female and male talkers,'' J. Acoust. Soc. Am. **87,** 820–857.

Kluender, K. R., Diehl, R. L., and Wright, B. A. (**1988**). ''Vowel-length differences before voiced and voiceless consonants: an auditory explanation,'' J. Phon. **16**, 153–169.

Li, X., and Pastore, R. E. (**1995**). ''Perceptual constancy of a global spectral property: Spectral slope discrimination,'' J. Acoust. Soc. Am. **98,** 1956–1968.

Lindau, M. (**1975**). ''Features for Vowels,'' Univ. California, Los Angeles, Working Papers in Phonetics, No. 30.

Lindau, M. (**1978**). ''Vowel features,'' Language **54**, 541–563.

Lindau, M. (**1979**). ''The feature 'expanded','' J. Phon. **7**, 163–176.

Local, J. K. (**1995**). ''Making sense of dynamic, non-segmental phonetics,'' in *Proceedings of the 13th International Congress of Phonetic Science*, edited by K. Elenius and P. Branderud, Vol. 3 (Stockholm), pp. 2–9.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge U.P., New York).

Macmillan, N. A., and Kingston, J. (**1995**). ''Integrality, separability, and configurality: The psychophysics of the Garner paradigm,'' in *Fechner Day 95*, edited by C.-A. Possamai (Intl. Soc. Psycho-Physics, Cassis, France), pp. 243–248.

Macmillan, N. A., and Kingston, J. (**submitted**). ''Integrality, separability, and configurality: The psychophysics of the Garner paradigm,'' Psychol. Rev. (submitted).

Maddieson, I., and Ladefoged, P. (**1985**). '' 'Tense' and 'lax' in four minority languages of China,'' J. Phon. **13**, 433–454.

Maddox, W. T. (**1992**). ''Perceptual and decisional separability,'' in *Multidimensional Models of Perception and Cognition*, edited by F. G. Ashby (Erlbaum, Hillsdale, NJ), pp. 147–180.

Marascuilo, L. A. (**1970**). ''Extensions of the significance test for one-parameter signal detection hypotheses,'' Psychometricka **35**, 237–243.

Nearey, T. M. (**1995**). ''A double-weak view of trading relations: Comments on Kingston and Diehl,'' in *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence*, edited by B. Connell and A. Arvaniti (Cambridge U.P., Cambridge, U.K.), pp. 28–40.

Ohala, J. J. (**1981**). ''The listener as a source of sound change,'' in *Papers from the Parasession on Language and Behavior*, edited by C. S. Masek, R. A. Hendrick, and M. F. Miller (Chicago Linguistic Society, Chicago), pp. 178–203.

Ohala, J. S. (**1985**). ''Around flat,'' in *Phonetic Linguistics*,'' edited by V. Fromkin (Academic, Orlando).

Perkell, J. S. (**1969**). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (MIT, Cambridge, MA).

Thorburn, R., Walsh, L. J., Macmillan, N. A., and Kingston, J. (**1994**). ''Components of integrality in the perception of voice quality and tongue root position,'' J. Acoust. Soc. Am. **95,** 2871(A).

Tucker, A. N. (**1966**). *Linguistic Analyses: The Non-Bantu Languages of North-East Africa* (Oxford U.P., London).

Walsh (Dickey), L., Bartels, C., Thorburn, R., Kingston, J., and Macmillan, N. A. (**1995**). ''Laxness integrates with F_1 (Usually, but not always, negatively),'' J. Acoust. Soc. Am. **97,** 3421(A).