

# Thresholds for second formant transitions in front vowels<sup>a)</sup>

Diane Kewley-Port<sup>b)</sup> and Shawn S. Goodman

*Department of Speech and Hearing Sciences, Indiana University, Bloomington, Indiana 47405*

(Received 22 August 2003; revised 3 August 2005; accepted 29 August 2005)

Formant dynamics in vowel nuclei contribute to vowel classification in English. This study examined listeners' ability to discriminate dynamic second formant transitions in synthetic high front vowels. Acoustic measurements were made from the nuclei (steady state and 20% and 80% of vowel duration) for the vowels /i, I, e, ε, æ/ spoken by a female in /bVd/ context. Three synthesis parameters were selected to yield twelve discrimination conditions: initial frequency value for F2 (2525, 2272, or 2068 Hz), slope direction (rising or falling), and duration (110 or 165 ms). F1 frequency was roved. In the standard stimuli, F0 and F1–F4 were steady state. In the comparison stimuli only F2 frequency varied linearly to reach a final frequency. Five listeners were tested under adaptive tracking to estimate the threshold for frequency extent, the minimal detectable difference in frequency between the initial and final F2 values, called  $\Delta F$  extent. Analysis showed that initial F2 frequency and direction of movement for some F2 frequencies contributed to significant differences in  $\Delta F$  extent. Results suggested that listeners attended to differences in the stimulus property of frequency extent (hertz), not formant slope (hertz/second). Formant extent thresholds were at least four times smaller than extents measured in the natural speech tokens, and 18 times smaller than for the diphthongized vowel /eI/. © 2005 Acoustical Society of America. [DOI: 10.1121/1.2074667]

PACS number(s): 43.71.Es, 43.66.Fe [PA]

Pages: 3252–3260

## I. INTRODUCTION

According to a traditional description of vowel sounds, the steady-state formant frequencies F1 and F2 have been considered the most important perceptual cues (Peterson and Barney, 1952). For over 20 years, however, research has shown that for American English (AE) dynamic formant information also plays an important role in vowel recognition (Strange, Verbrugge, Shankweiler, and Edman, 1976; Strange, 1989; Andruski and Nearey, 1992; Strange, Jenkins, and Johnson, 1993; Hillenbrand, Getty, Clark, and Wheeler, 1995; Hillenbrand and Nearey, 1999). Vowels, of course, usually occur in consonantal context, for example the common syllable type in AE, the CVC. Two kinds of voiced formant movement occur in CVCs: movement during the transition portion between consonants and vowels, and movement in the center portion (nucleus) of the vowel. Formant transitions carry acoustic cues for both consonants and vowels (Lieberman and Mattingly, 1985; Nearey, 1989; Kewley-Port, 1995; Ohde and Abou-Khalil, 2001). During the initial and final portions of the syllable, formant transitions are characterized by short duration and usually rapid movement across large frequency ranges. Vowel nucleus dynamics are characterized by longer durations and slower changes across smaller frequency ranges, most clearly seen in vowels like [e<sup>I</sup>] and [o<sup>U</sup>]. Fully diphthongized vowels such as /aI/ and /oi/ are characterized by long durations with extensive changes in frequency. Perhaps because of the more limited movement in non-diphthongized vowels, the nucleus

dynamics have often been characterized as 'quasi-steady state.' However, it has been shown that systematic movement does occur (Nearey, 1989; and Hillenbrand *et al.*, 1995).

The dynamic cues in syllables available in either the consonantal transitions alone (Strange, 1989) or the nucleus region alone (Nearey and Assmann, 1986) have been shown sufficient to produce high vowel identification rates (Hillenbrand *et al.*, 1995; Andruski and Nearey, 1992; Hillenbrand and Nearey, 1999). A complete theory of vowel perception must explain the roles of both types of formant movement in vowel identification. In fact, to model the underlying processing mechanisms of vowel perception, the ability to discriminate small differences in vowel dynamics should also be known. A reasonable hypothesis is that dynamic formant cues cannot be used for vowel identification if they cannot be perceived. Yet little is known about how much formant movement is needed to exceed thresholds for discriminating dynamic changes. The purpose of the present study is to estimate the psychophysical thresholds for discrimination of formant movement in the vowel nucleus. With few exceptions, previous threshold studies that used dynamic stimuli investigated transitions more similar to consonantal dynamics, i.e., transitions that are short and rapid. For instance, work has been done using short-duration tone glides (e.g., Summers and Leek, 1995; Madden and Fire, 1996, 1997; Sek and Moore, 1999) and rising or falling single-formant, speech like transitions (Porter, Cullen, Collins, and Jackson, 1991; van Wieringen and Pols, 1994). Porter *et al.* (1991) measured jnds for speech formant transitions. Stimuli consisted of a single formant modeled after F2 consonantal transitions connected to a steady-state portion. Because of the great temporal and spectral differences, interpretation of

<sup>a)</sup>Portions of this work were presented in a talk given at the 138th meeting of the Acoustical Society of America on June 4, 2000 in Atlanta, GA.

<sup>b)</sup>Electronic mail: kewley@indiana.edu

TABLE I. Average F1 and F2 formant values in hertz for the female talker for three repetitions of the words shown. Measurements were made at three intervals, 20%, the steady-state (SS, visually identified from spectrograms) and 80% of the total vowel duration.

	F1			F2		
	20%	SS	80%	20%	SS	80%
/bid/	319	303	320	2525	2600	2646
/bId/	358	363	391	2253	2251	2038
/bed/	486	452	347	2291	2438	2666
/bEd/	519	588	621	2060	2068	1893
/bæd/	684	698	784	2075	2115	1846

these results for nucleus dynamics is probably not predictive of thresholds of formant dynamics in vowel nuclei.

A related study has examined dynamics in nonspeech stimuli, namely, harmonic profiles (Drennan and Watson, 2001). Thresholds were estimated for profile stimuli that were harmonically spaced (200 Hz) when fundamental frequency was swept over time. The duration was similar to that of a long vowel nucleus (400 ms). No consistent differences in thresholds for static versus dynamic stimuli were found.

Previous studies in this series have examined thresholds for formant frequency in stimuli that represent steady-state vowel nuclei. Kewley-Port and Watson (1994) measured thresholds for frequency difference in steady-state first and second formants of isolated female vowels. Their stimuli were comparable in length to vowel nuclei (160 ms). In that study, thresholds for discrimination were constant over the F1 region at about 14 Hz, and constant in the F2 region with a Weber ratio  $\Delta F/F$  of 1.5%. Other studies (e.g., Sinnott and Kreiter, 1991; Hawks, 1994) have reported similar results. Kewley-Port and her colleagues have extended these studies to examine thresholds in CVC context, and in phrases and sentences. Together these studies have yielded systematic descriptions of the degrading effects of phonetic context and other experimental variables on the ability for humans to discriminate small differences in vowel formants. However, the vowel nuclei in all these studies used steady-state formants (i.e., frequency in the vowel nucleus did not change over time), while the formants in the nuclei of most natural AE vowels vary in frequency over time.

The present study was designed to examine the discrimination of formant transitions that more closely resemble natural AE vowel nuclei. An important issue for a theory of vowel perception is to determine the stimulus conditions in which nucleus dynamics are the information-bearing properties of vowels. If the observed vowel dynamics are not above threshold, then these differences cannot contribute information to vowel classification. Specifically, this study determined the smallest amount of frequency change in a second formant that could be discriminated under nearly optimal listening conditions by well-trained listeners. A working hypothesis for this study was that if formant dynamics in the nucleus contribute to vowel identification, they should exceed optimal thresholds for discriminating formant transitions by an amount sufficient to be salient in running speech, e.g., at least a factor of two. Because vowels are often modeled with steady-state formants, the standard vowels used as the basis for the formant transition comparisons were chosen

to have static (flat) formant frequencies. Synthesis parameters manipulated in the comparison stimuli, based on measurements of natural vowels from a female talker, included initial frequency of F2, slope direction, duration, and F1 frequency. This study determined thresholds using psychophysical procedures for dynamic changes in F2 frequency for five synthetic front vowels.

## II. METHOD

### A. Participants

Six young (under 35 years) listeners with normal hearing participated in this study. All listeners had audiometric thresholds of 15 dB HL or better at octave intervals from 250 to 4000 Hz. Listeners were paid for their participation. One listener could not produce consistent thresholds and voluntarily terminated participation. This study reports on data obtained from the remaining five listeners.

### B. Stimuli

To select the acoustic parameters to manipulate in these psychoacoustic studies of vowel nuclei, we derived the stimulus parameters from natural speech. The American English front vowels, /i, I, e, ε, æ/, recorded digitally at a 10 kHz sample rate, were analyzed. The vowels in the original utterances were produced by a female talker from the North Midland dialect region (Labov, 2004) in /bVd/ context in short sentences. Formants were measured for each word using linear predictive coding (LPC) analysis with a 128-point Hamming window, preemphasis, and 12 coefficients. To specify vowel formant change primarily in the nuclei and reduce the consonantal effects, formant values for F1 and F2 were measured at 20% and 80% of vowel duration, following Hillenbrand *et al.* (1995) for three repetitions of each vowel ( $n=15$ ). Also, the apparent F2 steady-state value was determined visually and measured from spectrograms. These measures are shown in Table I.

Based on these acoustic measurements, synthesis parameters were selected with the goal of producing vowels that resembled the intended vowels even though various simplifications were imposed. It was observed that the initial F2 values of /I/ and /e/ (at 20%) were very close in frequency (2253 and 2291 Hz), as were the initial F2 values of /æ/ and /ε/ (2060 and 2075 Hz). Thus to simplify the experiment, the mean value of each pair was used as the initial frequency value to represent both vowels. The measured initial F2 fre-

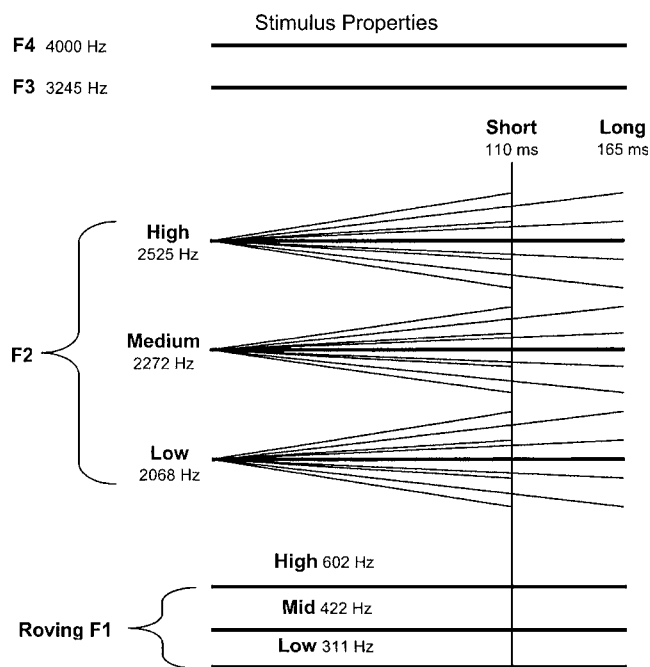


FIG. 1. Stylized formants represent the major stimulus variables for the synthetic vowels used in this study; three values of F1, three onset values of F2, rising and falling slope direction, and short or long duration.

quency of /i/ was substantially higher than the others. Thus, the resulting initial second formant frequency values were 2525 Hz for /i/, 2272 Hz for /I-e/, and 2068 Hz for /ε-æ/.

From the acoustic analysis, it was noted that across the nucleus portion, each vowel had either a general rise or fall in F2 frequency (Table I). Thus a second experimental variable of either a rising or a falling F2 slope was chosen. The set of vowels analyzed contained both long (/i, e, æ/) and short (/I, ε/) vowels. In order to examine possible effects of vowel length, two durations were calculated from the average of the long vowels (154, 166, 173 ms) and the average of the short vowels (100, 124 ms), rounded to 5 ms: long (165 ms) and short (110 ms). Together these three factors yielded 12 stimulus conditions: initial F2 frequency ( $\times 3$ ), direction of formant movement ( $\times 2$ ), and duration ( $\times 2$ ). Clearly some of the combinations do not occur in AE vowels, but the intention of this study was to systematically investigate the effects of all three factors. The general pattern of these 12 F2 transitions is shown in Fig. 1.

Normally a change in second formant frequency is coupled to a change in first formant frequency. F1 frequency was also of interest as a parameter, but in order to focus on the movement in F2 with simpler acoustic stimuli, it was decided to fix F1 as a steady-state parameter. Because the F1 values measured in the original 15 words ranged over 370 Hz, it appeared necessary to select several F1 values in order to have the stimuli resemble AE vowels. Three values of F1 were calculated using a similar grouping strategy as for F2 (/i/, /I-e/, and /ε-æ/). This resulted in three steady-state frequency values for F1: low (/i/, 311 Hz), medium (/I-e/, 422 Hz), and high (/ε-æ/, 602 Hz). Since the front vowels used in this experiment have a wide separation between F1 and F2, it was hypothesized that F1 frequency would not affect thresholds for F2 formant movement; therefore F1 frequency was

rovved randomly throughout the experiment. Steady-state values were chosen for F3 (3245 Hz) and F4 (4000 Hz). The bandwidths of the first three formants were 70, 90, and 170 Hz, for F1, F2, and F3, respectively.

The fundamental frequency was chosen to be steady state at 200 Hz (appropriate for this female talker). Although this choice also compromised the naturalness of the stimuli, a steady-state F0 was chosen to avoid possible undesirable interactions between shifting harmonics and their placement relative to the moving F2 formant peaks.

Using the above parameters, the vowel nuclei were synthesized with the cascade branch of the KLTSYN synthesizer (see Klatt, 1980). Stimuli had a length of either 165 or 110 ms with an additional 15 ms rise time and a 20 ms fall time. Eighteen *standard vowels* were synthesized with steady-state F1 and F2 values (3 F2 Frequencies  $\times$  2 lengths  $\times$  3 F1 frequencies). Of course these standard stimuli sounded unnatural because of the various constraints imposed to systematically manipulate the stimulus factors. Nonetheless, to document how these 18 synthetic stimuli would be identified by naïve American English listeners, a brief identification experiment was performed. Besides the 18 standard vowels, another set of stimuli was included that more carefully preserved the parameters measured from the recorded utterances. This set of five “original” stimuli were synthesized with both F1 and F2 formant frequencies linearly changing from the average 20% to the 80% intervals measured for each of the five natural syllables (shown in Table I), and with the appropriate long (165 ms) or short (110 ms) duration, although F0, F3, and F4 were still steady state. Eight naïve, normal-hearing American English listeners from Indiana were recruited for the identification task. The keyword responses included the ten monophthongal vowels in order to provide a wide choice of vowel responses. Following a short training task with another female speaker’s syllables, listeners heard six repetitions of the 28 vowels.

Overall subjects selected the five front vowels from the ten-vowel response set 78% of the time. Identification results showed that the original stimuli modeling the F1 and F2 transitions were correctly identified at 58% on average, with /i/ most accurately identified at 88%. The 18 standard vowels, all with fixed formants, had no predicted category labels. Two results were notable. The stimuli with the high F2 and low F1 were categorized 95% as /i/. Second, in all cases changes in F1 from low to mid to high values radically altered the perceived vowel quality. For example, for a stimulus with about 40% classification of a vowel with one value of F1, the percentage responses dropped to 15% and then 0% as F1 changed.

Summarizing, although the values for the parameters of F1, F2, vowel duration and formant movement were derived from natural speech, the resulting compromises in stimulus construction yielded stimuli that were not well categorized as AE vowels except for /i/. Clearly, however, the acoustic parameters manipulated in this study of thresholds for formant movement are information bearing properties of AE vowels.

Sets of test stimuli were also synthesized for the comparison conditions. Each test set had F2 values that either

increased or decreased in frequency over the entire stimulus. These changes in F2 were specified as the amount of frequency change in hertz between onset and offset (called *frequency extent*). The shape of the F2 transition was determined by linear interpolation calculated by KLTSYN between the initial and final values. Stimuli for each F2 transition were synthesized three times, once for each of the F1 frequencies. Thus there were 36 sets, of stimuli (3 F2 frequencies  $\times$  2 lengths  $\times$  3 F1 frequencies  $\times$  2 slope directions). For each set, the change in frequency of the F2 extent varied in 14 steps. Based on pilot work, unequal step sizes were used to change the frequency of F2 extent. The largest steps used increments of around 50 Hz, while the smaller steps had 5 Hz increments for rising F2s and 8 Hz decrements for falling F2s. During testing a few vowel stimuli proved more difficult than expected for some listeners, and to ensure that accurate thresholds were being measured, new sets were synthesized wherein the smaller steps had 7 Hz increments for rising F2s and 10 Hz decrements for falling F2s.

### C. Apparatus

Listeners were seated in a sound-treated booth facing a computer monitor and keyboard. Stimuli were presented monaurally to the right ear through a TDH-39 headphone. Up to three listeners were tested simultaneously, each listening to a separate output through a Tucker-Davis Technologies (TDT) array processor in a 486 computer. Stimuli were output through a 16-bit digital-analog (D/A) converter at a 10-kHz sample rate (TDT DA1) followed by a 4.3 kHz low-pass filter with a 96 dB/octave rolloff uploaded to a TDT PF1.

One standard vowel (medium F2=2272 Hz, medium F1=422 Hz) was selected as the calibration vowel. Output gain was set so the sound pressure level measured in an NBS-9A coupler with a Larson Davis sound level meter (model 800B) on linear weighting was 70 dB SPL. The levels of the other eight standard vowels [(3 values of F2  $\times$  3 values of F1)-1] were adjusted via the overall gain parameter of the KLTSYN synthesizer so that their overall rms energy was within  $\pm 1$  dB of the calibration stimulus. One standard could not meet this tolerance level without undergoing severe peak clipping in the synthesis. Its level was accordingly adjusted to within  $\pm 2$  dB of the calibration stimulus.

### D. Procedure

A two-down-one-up adaptive tracking procedure (Levitt, 1971) estimated the difference in F2 offset frequency between standard and comparison stimuli at 70.7% correct. On each trial, listeners heard three successive stimuli. The standard stimulus was always presented first, followed by presentation of the standard and comparison stimuli in random order. Listeners indicated which of the last two intervals contained the stimulus that sounded different from the original reference by pressing the appropriate key on a keyboard. Feedback for correct responses was provided following each trial. Each block consisted of 90 trials. To obtain nearly op-

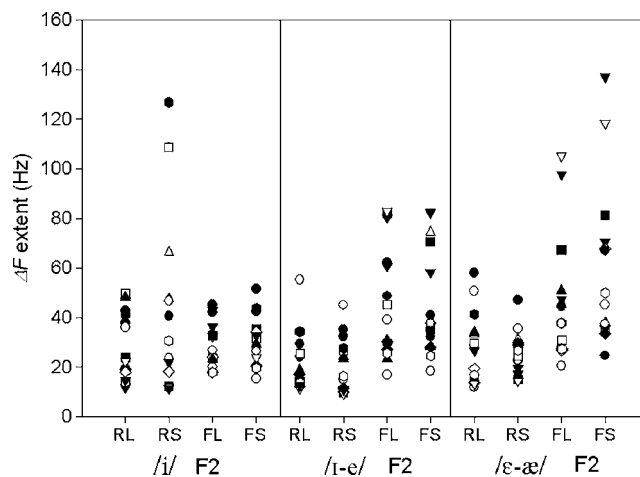


FIG. 2. Each symbol is a threshold estimated for  $\Delta F$  extent for one listener for all 180 data points [F1(3)  $\times$  F2(12)  $\times$  listeners(5)]. Panels are organized by the onset of F2 frequency, and within panels for rising long (RL), rising short (RS), falling long (FL), and falling short (FS) formants for a total of 12 stimulus conditions.

timal thresholds, only one of the 12 stimulus conditions for F2 was presented per block, although the three F1 frequencies were roved randomly within a block.

In order to obtain reliable thresholds, listeners were highly trained. Listeners were trained with the parameters of the stimuli set to medium F2, long duration, and rising slope. This condition was chosen as the training stimulus because rising slopes were easier than the falling slopes in pilot work. During training, listeners completed an estimated 3500 trials each over four sessions.

Following training, listeners were tested five days per week and were allowed to proceed at their own pace. A test session lasted for approximately 1.5 h per day, including breaks. On average, listeners completed 11 blocks of 90 trials each day. The 12 stimulus conditions were tested in random order across subjects. For each condition, blocks were run until it was judged that stable values had been achieved based on examination of block-by-block plots of the reversal means. The stopping rule was near-asymptotic performance over four blocks. Wide variability was seen in the amount of time spent on each condition. The average time spent on each condition was 14.43 blocks (1299 trials), but ranged from 6 to 30 blocks (540 to 2700 trials). No clear patterns emerged to explain this variability.

For each block, a value  $\Delta F$  in hertz was calculated as the difference in the formant frequency extent between the standard and test stimuli, using the average of the reversals, excluding the first three reversals. A threshold for formant frequency extent,  $\Delta F$  extent, for each listener was averaged over the last four blocks.

### III. RESULTS

Results, reported as  $\Delta F$  extent, were first examined for individual variability. Figure 2 displays the 180 independently measured data points for the 12 F2 conditions, each with three F1 formants, from the five listeners (15 points/condition). As can be seen in Fig. 2,  $\Delta F$  extent values had narrow distributions for seven of the 12 conditions while in

TABLE II.  $\Delta F$  extent thresholds in hertz for F2 transitions for the four factors of initial F2 frequency (low, mid, high), vowel length (short, long), slope direction (rising, falling), and fixed value of F1 (low, mid, high). The F1 values were calculated as arithmetic averages over the five subjects. The pooled data across the five subjects and three values of F1 ( $N=15$ ) were calculated from the antilog of the means of the  $\log_{10}$  thresholds (see text). All values in hertz.

Slope	F1	F2 Short			F2 Long		
		/ $\varepsilon$ - $\alpha$ /	/I-e/	/i/	/ $\varepsilon$ - $\alpha$ /	/I-e/	/i/
		2068	2272	2525	2068	2272	2525
Rising	Low=311	31.7	26.4	51.3	38.7	29.1	28.7
	Mid=422	25.3	16.2	40.9	22.0	17.0	31.6
	High=602	21.8	19.0	46.6	21.1	16.7	29.4
	Mean (antilog)	19.6	15.8	38.6	18.8	16.3	23.6
Falling	Low=311	65.7	44.1	34.8	49.0	40.9	31.9
	Mid=422	58.4	42.3	27.8	48.2	41.7	24.3
	High=602	51.9	50.8	26.9	41.7	54.4	32.6
	Mean (antilog)	45.6	44.2	25.0	38.1	44.0	30.0

the other five conditions there was a positive skew in the distributions. This presence of a few highly elevated thresholds is similar to that reported for other complex stimuli, for example, vowel formant thresholds (Kewley-Port and Watson, 1994) and dynamic profiles (Drennan and Watson, 2001). However, this resulted in a high correlation ( $r=0.80$ ) between the means and the variances of the data. The data were therefore subjected to a  $\log_{10}$  transform. This transform created a more normal distribution and reduced the correlation between means and variances ( $r=0.26$ ). Thus statistical analyses were carried out on the  $\log_{10}$  transform of the 180 individual data points. Average thresholds are reported in hertz (e.g., Table II and in subsequent figures) by taking the antilogs of the transformed values, an averaging method that dampens the effect of the skewed thresholds. A Mauchly Sphericity test was also performed on the data, and results suggested that the assumption of sphericity was violated (Max and Onghena, 1999). A multivariate analysis of variance (MANOVA) was therefore preferred to the univariate three-way repeated measures ANOVA. The results of the MANOVA may be interpreted in much the same way as the results of an ANOVA.

A four-factor MANOVA was calculated for F1 frequency, F2 frequency, slope direction, and length, treated as within-subject factors, and  $\log_{10} \Delta F$  extent as the dependent variable. The  $\Delta F$  extent thresholds for these four factors, and their means, are shown in Table II. Results showed a significant main effect only for F2 frequency [ $F(2,3)=14.96$ ,  $p < 0.028$ ]. A significant two-way interaction was obtained for F2 frequency  $\times$  slope [ $F(2,3)=22.36$ ,  $p < 0.016$ ]. There was also one significant three-way interaction for F1 frequency  $\times$  F2 frequency  $\times$  length [ $F(4,1)=11\ 818.15$ ,  $p < 0.007$ ].

### A. Effect of F1

In this experiment it was not anticipated that F1 frequency would affect discrimination of F2 transitions. Indeed, no main effect was found for F1, or for any two-way interactions involving F1. The values of  $\Delta F$  extent for F1 averaged over listeners are shown in Table II. Given the significant three-way F1  $\times$  F2  $\times$  length interaction, results were

scrutinized for specific F1 effects. The largest effects were for long stimuli when F1 was low and for F2 for /i/ or / $\varepsilon$ - $\alpha$ / stimuli. Overall the effect of F1 on F2 transition thresholds was small and not systematic.

### B. Effect of F2 frequency and slope direction

The significant main effect for F2 frequency reflected that  $\Delta F$  extent was smaller for the /i/ and /I-e/ frequencies compared to the / $\varepsilon$ - $\alpha$ / F2 frequency. Although there was no effect of the F2 slope direction overall [ $F(2,3)=3.94$ ,  $p > 0.10$ ], the significant interaction for F2 frequency  $\times$  slope direction reflected that  $\Delta F$  extent was smaller for rising than falling second formants for /I-e/ F2 and / $\varepsilon$ - $\alpha$ / F2 as shown in Fig. 3. For /i/ F2, however,  $\Delta F$  extent was nearly the same for rising and falling. Given these results were based on the  $\log_{10}$  transformed data, additional analyses of the data were conducted to verify that the above interpretation was correct. For example, the medians of the  $\Delta F$  extent values in hertz represented in Fig. 3 produced a display with the same pattern of interactions.

Note that in Fig. 3 it appears that thresholds for falling F2 transitions increase as frequency decreases. It is unlikely

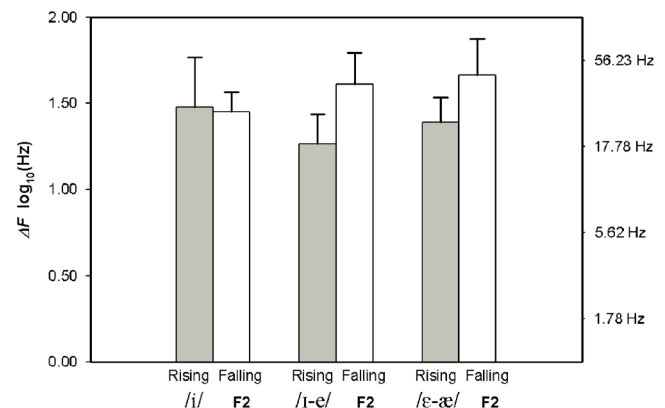


FIG. 3.  $\Delta F$  extent thresholds calculated from the  $\log_{10}$  transformed data to show the interaction between the F2 onset frequencies and slope direction. Error bars indicate standard deviation in the  $\log_{10}$  units. The right-hand axis indicates equivalent values in hertz.

that the increase in threshold for falling slopes was due to a closer proximity of F1 than for rising slopes, because the highest F1 center frequency was 602 Hz and the lowest F2 center frequency was 2011 Hz. Thus there was always at least 1329 Hz separating the F1 and F2 formants for the vowels investigated here.

### C. Effects of duration

There was no significant difference between thresholds for the long and short durations. Thus slope is not the primary acoustic property used in discrimination, because the slopes for the short conditions are markedly steeper than the slopes for the long conditions (0.167 versus 0.280, 0.167 versus 0.247, 0.191 versus 0.328 Hz/ms for increasing F2). Apparently listeners pay attention to the extent of formant frequency movement (hertz) rather than the slope of formant movement (hertz/milliseconds), a conclusion compatible with those of Nearey and Assmann (1986).

## IV. DISCUSSION

### A. Formant transition thresholds

The purpose of this study was to determine thresholds for discriminating dynamic changes in the second formant frequency of front vowels. Four acoustic properties that contribute to the accurate identification of American English vowels were manipulated, F1 frequency, F2 frequency, slope direction, and length. Thresholds differed primarily as a result of manipulating onset of F2 frequency and for specific combinations of F2 onset with the slope of the formant transitions. Given our psychophysical approach to investigating these formant transitions, it is of theoretical interest to know how auditory sensitivity for frequency change in formants compares to that for simple tones (tone glides). The most comparable research is that of Dooley and Moore (1988). In the Dooley and Moore (1988) second experiment, tone glides started at 2000 Hz, with durations of either 100 or 200 ms long. Similar to our formants, duration had little effect on thresholds for either set of stimuli. They obtained thresholds around 15 Hz in comparison to our F2 thresholds ranging from 19–52 Hz. Thus the lowest thresholds in Fig. 3 for the rising /I-e/ F2 were similar in magnitude to those for tone glides, while the highest F2 thresholds were about 300% higher than for tones. These threshold differences reflect our significant F2 X slope interaction. However, slope direction appears to affect dynamic formants differently than tone glides because slope direction had no effect in Dooley and Moore (1988) for their shorter tone glides in experiment 2. However, their results for longer glides are the opposite of ours, with thresholds being smaller for falling tone glides than for rising glides.

Summarizing, thresholds for our F2 transitions were roughly similar to those for tone glides with similar parameters, except that thresholds for falling formant transitions increased as F2 frequency decreased. While it is unclear what reasons might underlie such differences in the effect of slope direction for simpler versus more complex vowel

stimuli, we suspect that the presence of the fixed formants (F1, F3, and F4) are an important factor, as well as a possible interaction between F0 and F2.

How do formant thresholds patterned after vowel nuclei compare to those patterned after stop consonants? The most comparable data for stop transitions is that of Porter *et al.* (1991). They used a single formant transition rising or falling to 1800 Hz followed by about a 200 ms steady-state segment. For their 120 ms transition, the thresholds range from 50 to 150 Hz. These single formant thresholds are about 300% greater than for ours for vowel formant stimuli. Of the many differences between experiments, one reason for the higher thresholds may be that listeners in Porter *et al.* (1991) participated for 1–2 h, while our listeners were trained extensively before data collection started (5 h training, and an average of 24 h testing). Another difference between vowel and stop transitions is that thresholds were also smaller for falling single formants than rising ones in Porter *et al.* (1991). Thus auditory processing of stop-like, single formant stimuli appears to be different from that of vowel-like, multiformant stimuli.

An important question for understanding the processing of dynamic stimuli is whether transition slope or frequency extent provides the information to differentiate frequency transitions. For our vowel transition stimuli, the short (110 ms) versus long (165 ms) duration did not have an effect on the  $\Delta F$  extent thresholds, and therefore frequency extent was the distinguishing acoustic property. These results are similar to those reported for tone glides by both Madden and Fire (1997) and Moore and Sek (1998) who found no threshold difference between the 50 and 400 ms stimuli. An earlier study by Nábělek and Hirsh (1969) also found that  $\Delta F$  extent was an important acoustic property in the perception of tone glides. Results are not as clear for single formant studies. While Porter *et al.* (1991) and van Wieringen and Pols (1994) reported that thresholds increased significantly as the single formants shortened, additional experiments by van Wieringen and Pols (1994) did not reveal whether  $\Delta F$  extent or transition rate was the primary acoustic property for these short stimuli. Thus it appears that for several different types of dynamic stimuli including vowels, the perceptually salient acoustic property is the frequency achieved at the end of the transition,  $\Delta F$  extent, not transition slope or rate.

### B. Comparison to steady-state formants

Given that most sounds are dynamic, but most frequency thresholds are measured from static stimuli, it is important to understand differences in auditory processing of dynamic versus static stimuli. The present thresholds for dynamic formants can be directly compared to static vowel thresholds reported by Kewley-Port and Watson (1994). They used F2 frequency values of 2900, 2500, and 1950 Hz to represent the vowels /i, e, æ/ measured from the same talker used here. Using experimental procedures similar to those of the present study, they estimated  $\Delta F$ , the minimum frequency change in a steady-state F2 needed to detect either an increment or decrement in frequency from a steady-state standard. Vowel duration for the steady-state stimuli was 160 ms com-

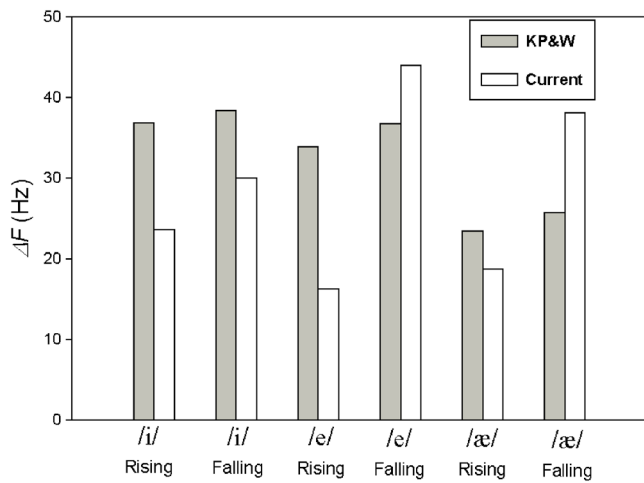


FIG. 4. F2 thresholds from two studies are displayed. The light grey bars are  $\Delta F$  thresholds for steady-state formants from Kewley-Port and Watson (1994). The white bars are  $\Delta F$  extent thresholds from the current study. Note that the onset frequencies are somewhat different between the two studies. The labels rising or falling indicate a positive or negative frequency shift for the steady-state formants, and the slope direction of the formant transitions.

pared to 165 ms in the present study. Thresholds are plotted for both experiments in Fig. 4 for both the rising and falling changes in frequency.

For the six formant comparisons, the differences between  $\Delta F$  extent thresholds and steady-state thresholds ranged from  $-14.2$  to  $16.1$  Hz. While this variability is somewhat high, the overall means of the thresholds were similar [ $30.9$  versus  $32.5$  Hz for dynamic versus steady-state,  $t(10) = 0.36$ ,  $p = 0.73$ , two-tailed]. Moreover, no general pattern of differences in thresholds in Fig. 4 between the dynamic and steady-state F2s is apparent. These results are similar to those for another type of complex stimuli, harmonic profiles. Drennan and Watson (2001) also observed no significant pattern of differences in amplitude detection thresholds between static and dynamic harmonic profiles. However, these results for harmonic complexes contrast with those for single sinusoids where thresholds for glides were significantly greater than for static tones (Dooley and Moore, 1988). Summarizing, although auditory processing appears to be influenced by several factors for harmonic complexes, static versus dynamic stimulus parameters do not produce consistent effects.

### C. Comparison to natural vowel transitions

Thresholds for formant transitions in this study were obtained under nearly optimal listening conditions. In natural speech it must be assumed that formant movement substantially exceeds optimal perceptual thresholds in order for communication to be robust in ordinary listening conditions. To quantify this relationship, the  $\Delta F$  extent thresholds were compared to natural F2 transitions measured from the three original recordings of /bVd/ stimuli (Table I). Of the 12 conditions tested, five correspond to the natural vowels /i, I, e, ε, æ/ measured from the original talker when the factors of F2 initial frequency, vowel duration, and F2 slope are matched. The length of the black bars in Fig. 5 represents the  $\Delta F$  extent thresholds, and the measured change in F2 from 20–80% of vowel duration is shown by hatched lines.

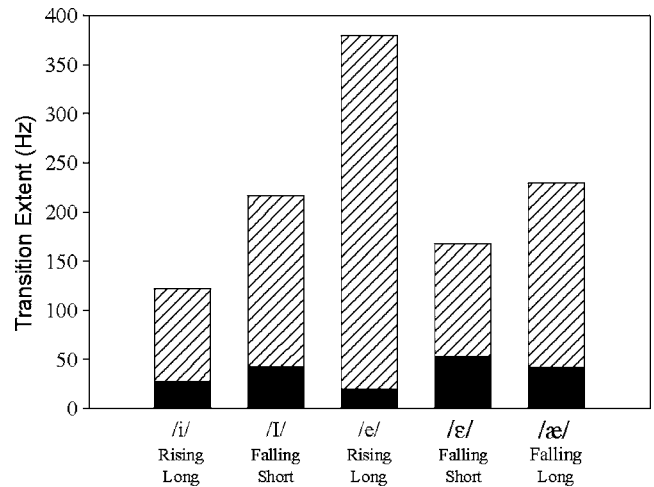


FIG. 5. The height of the black bars displays the  $\Delta F$  extent thresholds in this study for the stimulus conditions as labeled. The length of the hatched bars displays the average F2 transition extent measured from 20% to 80% of the vowel duration for the female talker's vowels recorded for this study.

Thresholds for  $\Delta F$  extent were rather constant and small relative to the amount of naturally occurring movement. The amount of naturally occurring movement was on average more than four times larger than the thresholds for  $\Delta F$  extent for the vowels /i, I, e, ε, æ/. For the highly diphthongized vowel /e/, the amount of naturally occurring movement was 18 times larger than the threshold for  $\Delta F$  extent. One reason for the large increase in the /e/ ratio was that the  $\Delta F$  extent threshold was smallest for /e/. Clearly naturally occurring vowel F2 transitions are perceptually very salient, given that the thresholds are better by a factor of at least four than the actually observed vowel transitions.

Although the present study examined only a small number of F2 transitions from a female talker's vowels, an intriguing hypothesis is suggested by the above analysis. It is clear from numerous reports that dynamic formants improve accurate vowel identification (c.f. Hillenbrand and Gayvert, 1993). In fact, Hillenbrand and Nearey (1999) showed that larger formant dynamics in their synthetic vowels correlated with more accurate vowel identification. Presumably, when formant dynamics are particularly important cues to vowel identity, a corresponding difference in the sensory abilities to process formant transitions should be observed. In the present data the best sensitivity to F2 transitions was for /e/, both in terms of the absolute lowest threshold and in relation to the amount of naturally occurring F2 transition (Fig. 5). A plausible hypothesis for the large differences in sensitivity is that the auditory system becomes more sensitive to formant dynamics for just the vowels, such as /e/, where formant change is a more important cue to vowel identity. This hypothesis is in agreement with one suggested by Guenther, Nieto-Castanon, Ghosh and Tourville (2004), namely, that the brain puts more neural resources into regions of the acoustic space where differences in sound are behaviorally more important. Additional data on formant transitions are needed to determine whether this hypothesis generalizes to all American English naturally produced vowels.

Another issue examined in this study was the effect F1 had on F2. In our perceptual study, there was no systematic

effect of F1 manipulation on the discrimination of F2 vowel transitions. Given that F1 and F2 onsets were widely separated by more than 1300 Hz, this was the expected result. However, F1 differences alone can affect vowel categorization. The classification of some American English vowels appears to rely primarily on differences in F1 when F2 values are very similar (Hillenbrand *et al.*, 1995). In a vowel perception study, Johnson, Fleming, and Wright (1993) reported that when F2 is fixed, listeners selected different vowel categories when F1 was manipulated. Our small identification experiment also had stimuli in which only F1 varied, and large changes in vowel identification were obtained. Thus we expect to observe significant interactions between F1 and F2 in formant discrimination for some other vowels, particularly for the back vowels where formants are close together such that changes in one formant clearly affect the amplitude and shape of the other.

#### D. Relation between formant extent and formant transition thresholds

The concept of a behavioral threshold implies that differences between a standard and its comparison are all equally detectable by the listener. Thus spectral differences between the flat formant standard and the just discriminable formant transition at threshold across formants are in some sense perceptually equivalent. One possibility is that the actual acoustic differences are the same, i.e., formant extent thresholds in hertz were constant for different formants. Clearly this was not true because variability in the formant extent thresholds exceeded 280% over the 12 conditions, ranging from 15.8 to 45.6 (Table II). Formant extent is a value in hertz based on the difference in *synthesis parameters*, offset minus onset of F2. The actual acoustic differences among stimuli resulting from the specified parameters still need to be measured. Note that the F0 for all vowel stimuli was constant at 200 Hz so that the frequencies of the harmonics for each stimulus were the same, namely, the 25 harmonics from 200 to 5000 Hz. Thus the primary physical differences between the standard and comparison stimuli were intensity level differences for the 25 harmonics. Although these intensity differences can be fit with a formant resonance (e.g., using LPC), the present analysis focuses on the intensity differences. To measure the harmonic component level differences, a discrete Fourier transform (DFT) was taken of each pitch pulse of the stimuli. For each harmonic component, the DFT of the standard was subtracted from the DFT of the comparison stimulus nearest threshold, as shown Fig. 6 for / $\epsilon$ - $\text{\ae}$ / F2 long, rising transition. The results showed a linear increase or decrease over time in the decibel level, primarily for the two harmonics closest to the formant peak.

To quantify the spectral differences between the standard and comparison at threshold, two metrics were calculated. They were the intensity differences in decibels for the harmonics with the maximum decrement and the maximum increment at the stimulus offset. Given that intensity changed linearly over time, any measures at other proportional intervals (i.e., 1/2, 1/3 etc) of the total duration would yield the same relative metrics. Across all 12 conditions at offset, the

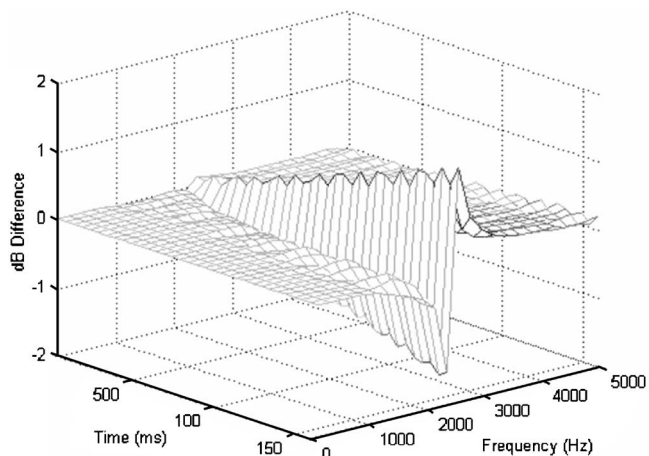


FIG. 6. For the / $\epsilon$ - $\text{\ae}$ / F2 long, rising stimulus, the acoustic difference between the DFTs of the standard and the comparison vowel nearest to threshold is displayed (see text).

decrement metric ranged from 1.53 to 2.69 dB, while the increment metric ranged from 1.5 to 3.58 dB. Not only are these two metrics far from constant, each correlated highly with  $\Delta F$  extent, namely,  $r=0.71$  for the increment metric and  $r=0.48$  for the decrement metric (both  $N=12$ ). Examining these metrics and other spectral-temporal measures of the transitions, we concluded that stimulus differences derived from acoustic measures could not yield a constant metric that captured the perceptual equivalence of the transition thresholds. Rather, some processing of this information in the auditory system must transform the physical differences into some internal representation that is more constant.

Obvious choices of models for auditory processing are the excitation and loudness pattern models developed by (Glasberg and Moore, 1990). We have previously been successful in applying them to find a constant metric for formant threshold data for steady-state vowels (Kewley-Port and Zheng, 1998; Liu, 2002). However, there appear to be no guidelines or logical extensions on how to apply these models to our stimuli with dynamic formants, and our modest efforts to apply these models to  $\Delta F$  extent thresholds were not successful. We note the application of excitation pattern models to tone glide discrimination (e.g., Moore and Sek, 1998) focused on changes in frequency while our stimuli had fixed-frequency harmonics with variable level differences. Thus, it appears that it is premature to adapt auditory models developed for frequency glides of sinusoids to the harmonic changes in formant transitions.

#### V. CONCLUSIONS

Listeners' ability to discriminate F2 transitions that are found in the nucleus of front vowels were reported for the manipulation of four factors, F1 frequency, initial F2 frequency, rising versus falling slope direction, and short versus long vowels in synthetic speech. The major results are summarized as follows:

- Changes in F2 onset affected  $\Delta F$  extent thresholds, but changes in F1 frequency, F2 slope direction, and length did not. An interaction between F2 onset and slope direction was seen for some conditions.
- Listeners attended to differences in the stimulus property of the frequency extent at the offset of the transition, not the formant slope.
- $\Delta F$  thresholds for vowel transitions were roughly similar to those for steady-state formants.
- Estimated  $\Delta F$  extent thresholds in hertz were similar to those obtained for tone glides with analogous acoustic parameters.
- The effects of several stimulus factors, e.g., rising versus falling slope, on  $\Delta F$  extent thresholds were the opposite of one another for vowel transitions versus tone glides, suggesting that the underlying perceptual mechanisms for discriminating frequency change are different for harmonic versus simple stimuli.
- Vowel transitions measured in natural speech exceeded the  $\Delta F$  extent thresholds in synthetic speech by at least a factor of four, indicating that formant dynamics observed in American English vowel nuclei are perceptually very salient.

## ACKNOWLEDGMENTS

This research was supported by NIHDCD-02229.

- Andruski, J. E., and Nearey, T. M. (1992). "On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables," *J. Acoust. Soc. Am.* **91**, 390–410.
- Drennan, W. R., and Watson, C. S. (2001). "Sources of variation in profile analysis. II. Component spacing, dynamic changes, and roving level," *J. Acoust. Soc. Am.* **110**, 2498–2504.
- Dooley, G. J., and Moore, B. C. (1988). "Duration discrimination of steady and gliding tones: A new method for estimating sensitivity to rate of change," *J. Acoust. Soc. Am.* **84**, 1332–1337.
- Glasberg, B., and Moore, B. C. J. (1990). "Deviation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., and Tourville, J. A. (2004). "Representations of sound categories in auditory cortical maps," *J. Speech Lang. Hear. Res.* **47**, 46–57.
- Hawks, J. W. (1994). "Difference limens for formant patterns of vowel sounds," *J. Acoust. Soc. Am.* **95**, 1074–1084.
- Hillenbrand, J., and Gayvert, R. T. (1993). "Vowel classification based on fundamental frequency and formant frequencies," *J. Speech Hear. Res.* **36**, 694–700.
- Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.* **97**, 3099–3111.
- Hillenbrand, J. M., and Nearey, T. M. (1999). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," *J. Acoust. Soc. Am.* **105**, 3509–3523.
- Johnson, K., Flemming, E., and Wright, R. (1993). "The hyperspace effect: Phonetic targets are hyperarticulated," *Lang.* **69**, 505–528.
- Klatt, D. (1980). "Software for cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Kewley-Port, D. (1995). "Thresholds for formant-frequency discrimination of vowels in consonantal context," *J. Acoust. Soc. Am.* **97**, 3139–3146.
- Kewley-Port, D., and Watson, C. S. (1994). "Formant-frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.* **95**, 485–496.
- Kewley-Port, D., and Zheng, Y. (1998). "Modeling formant frequency discrimination for isolated vowels," *J. Acoust. Soc. Am.* **103**, 1654–1666.
- Labov, W. (2005). *Atlas of North American English* (Mouton de Gruyter, New York) in press.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- Liberman, A. M., and Mattingly, I. G. (1985). "The motor theory of speech perception revised," *Cognition* **21**, 1–36.
- Liu, C. (2002). "Modeling vowel discrimination in quiet and noise for natural speech," unpublished doctoral dissertation, Indiana University.
- Madden, J. P., and Fire, K. M. (1996). "Detection and discrimination of gliding tones as a function of frequency transition and center frequency," *J. Acoust. Soc. Am.* **100**, 3754–3760.
- Madden, J. P., and Fire, K. M. (1997). "Detection and discrimination of frequency glides as a function of direction, duration, frequency span, and center frequency," *J. Acoust. Soc. Am.* **102**, 2920–2924.
- Max, L., and Onghena, P. (1999). "Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language and hearing research," *J. Speech Lang. Hear. Res.* **42**, 261–270.
- Moore, B. C., and Sek, A. (1998). "Discrimination of frequency glides with superimposed random glides in level," *J. Acoust. Soc. Am.* **104**, 411–421.
- Nábělek, I. V., and Hirsh, I. J. (1969). "On the discrimination of frequency transitions," *J. Acoust. Soc. Am.* **90**, 1510–1519.
- Nearey, T. M. (1989). "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.* **85**, 2088–2113.
- Nearey, T., and Assmann, P. (1986). "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297–1308.
- Ohde, R. N., and Abou-Khalil, R. (2001). "Age differences for stop-consonant and vowel perception in adults," *J. Acoust. Soc. Am.* **110**, 2156–2166.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Porter, R. J., Cullen, J. K., Collins, M. J., and Jackson, D. F. (1991). "Discrimination of formant transition onset frequency: Psychoacoustic cues at short, moderate, and long durations," *J. Acoust. Soc. Am.* **90**, 1298–1308.
- Sek, A., and Moore, B. C. (1999). "Discrimination of frequency steps linked by glides of various durations," *J. Acoust. Soc. Am.* **106**, 351–359.
- Sinnott, J. H., and Kreiter, N. A. (1991). "Differential sensitivity to vowel continua in Old World monkeys (*Macaca*) and humans," *J. Acoust. Soc. Am.* **89**, 2421–2429.
- Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.* **85**, 2135–2153.
- Strange, W., Jenkins, J., and Johnson, T. (1993). "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.* **74**, 695–705.
- Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (1976). "Consonant environment specifies vowel identity," *J. Acoust. Soc. Am.* **60**, 213–224.
- Summers, V., and Leek, M. R. (1995). "Frequency glide discrimination in the F2 region by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **97**, 3825–3832.
- van Wieringen, A., and Pols, L. C. (1994). "Frequency and duration discrimination of short first-formant speechlike transitions," *J. Acoust. Soc. Am.* **95**, 502–511.