Listener sensitivity to variations in the relative amplitude of vowel formants

Ewa Jacewicz
Department of Speech and Hearing Science, The Ohio State University, 1070 Carmack Road., Columbus, OH 43210
jacewicz.1@osu.edu

Abstract: Listener sensitivity to formant amplitude variations in the perception of both within- and between-vowel-category differences was examined for two isolated synthetic vowels, /i/ and /ɪ/. The relative amplitudes of *F*2 and *F*4 were varied in steps along acoustic continua which were created in accord with formant amplitude patterns found in natural speech. The results showed that vowel-specific patterns of formant amplitude variation affected listeners' judgments of "naturalness" of each vowel token, as well as its phonetic quality.

PACS numbers: 43.71. Es, 43.71.Bp, 43.66.Lj, 43.72.Ar
Date Received: November 2, 2004 **Date Accepted:** May 2, 2005

1. Introduction

In speech perception research, formant frequencies are generally assumed to be the principal determinants of vowel identity. The secondary spectral features such as formant bandwidth, spectral tilt, or formant amplitude relations are not regarded as essential to preserving vowel identity (Klatt, 1982). Although recognized, these secondary spectral characteristics which covary with vowel quality have not been incorporated with rigorous precision in the current models of vowel perception (e.g., Ito *et al.*, 2001; Hillenbrand and Houde, 2003). This paper attempts to gain more insight into the potential role of formant amplitude relations in the perception of both within- and between-category differences. A better understanding of the role of phonetic detail such as systematic amplitude variations examined here is a focus of current theoretical and experimental considerations (e.g., Hawkins, 2003).

Experimental manipulations of the relative amplitude of individual formants have shown that adjustments made to formant amplitudes can change the perceived vowel quality (e.g., Miller, 1953; Lindqvist and Pauli, 1968; Carlson *et al.*, 1970; Ainsworth and Millar, 1972; Aaltonen, 1985). As a general limitation of these studies, adjustments to formant amplitudes were made in a relatively arbitrary fashion, with little or no reference to the relations that may actually hold in naturally produced vowels. Addressing this shortcoming, two experiments were conducted in which such adjustments were guided by the patterns found in natural speech. Identifying the scope and range of perceptually detectable variations in formant amplitudes could lead to a better understanding of formant amplitude relations and to improved models of vowel recognition.

2. Experiment 1

Experiment 1 examined listener sensitivity to formant amplitude variations in the perception of within-category differences, which may contribute to perception of the vowel as a more or less natural-sounding exemplar of a category.

2.1 Stimuli

Prior to creating synthetic stimuli, acoustic analysis of natural vowels /i, ɪ/ was conducted. The corpus used was the Consonant–Nucleus–Consonant (CNC) monosyllabic word lists (Peterson and Lehiste, 1962), commercially recorded by a male speaker of American English on an audiotape for use in audiological testing. To analyze formant amplitude variations across a wide range of consonantal contexts, the selected words for /i/ were peak, beat, deep, deal, feet, sheep,

Table. 1. Mean formant amplitude (in dB) for the first four formants (*A*1,*A*2,*A*3,*A*4) for /i/ and /ɪ/ in natural speech. Also shown are the mean amplitudes relative to *A*1 (Rel*A*2,Rel*A*3,Rel*A*4). One standard deviation is shown in parentheses.

Vowel	*A*1	*A*2	*A*3	*A*4	Rel*A*2	Rel*A*3	Rel*A*4
/i/	64.0 (4.1)	41.6 (3.2)	40.5 (3.7)	41.2 (3.0)	−22.4 (3.1)	−23.5 (3.7)	−22.8 (5.6)
/ɪ/	66.4 (3.6)	52.3 (4.1)	43.4 (3.2)	36.3 (4.3)	−14.1 (4.8)	−23.0 (4.8)	−30.1 (4.5)

heat, reap, wheat, weak, neat, need, and the words for /ɪ/ were pick, bit, tip, dip, till, ship, hit, rib, wit, wig, knit. After digitizing the tokens at 22.05-kHz sampling rate, the vowels were analyzed for amplitude and frequency of the first four formants (autocorrelation LPC: 22.05-kHz sampling rate, 512-point Hamming window, 16 coefficients, 60% overlap) at the 25-ms interval where the energy of the vowel was the highest (Lienard and Di Benedetto, 1999). No pre-emphasis was applied for amplitude measurements. The amplitude values (in dB) were first measured using the reference level of a speech analysis program (TFR, 1999), which is the lowest nonzero instantaneous amplitude that is coded by the A/D conversion. Table 1 lists mean formant amplitudes (*A*1,*A*2,*A*3,*A*4) averaged across the consonant contexts for each vowel. Formant amplitude differences between /i/ and /ɪ/ were greatest for *A*2 and *A*4. These differences became even more apparent when the mean amplitude values were subsequently calculated relative to *A*1 for individual vowels (Rel*A*2,Rel*A*3,Rel*A*4). Comparing the values of Rel*A*2 and Rel*A*4, *F*2 is clearly stronger and *F*4 weaker for /ɪ/ than for /i/, whose Rel*A*2, Rel*A*3, and Rel*A*4 tend to be uniform. These differences suggest a possible contribution of *A*2 and *A*4 to the perceptual distinction between the two vowels. To test this, the levels of *A*2 and *A*4 were selected for systematic adjustment.

Four synthetic 4-formant series were created, in which either *A*2 or *A*4 varied in steps. The synthesis program used was a version of KLSYN (Johnson and Qi, 1987) in a parallel configuration with a 10-kHz sampling rate and 1-ms frame. The fixed frequency values for the first four formants for /i/ were 310, 2030, 2970, and 3400 Hz, and for /ɪ/ were 400, 1780, 2578, and 3400 Hz. The other parameters held constant for both vowels were f_0 (120 Hz), bandwidth for all four formants (50, 150, 300, and 400 Hz, respectively), and duration (88 ms). The following amplitude modifications were made. For the vowel /i/, the starting amplitude values for *A*1, *A*2, *A*3, and *A*4 were 60 dB, as suggested by Klatt. In the first 9-step continuum, only *A*2 was decremented in 2-dB steps beginning from 60 dB, while *A*1, *A*3, and *A*4 remained constant (compare Table 2). In the second 9-step continuum, *A*4 changed in decrements of 2 dB from the 60-dB starting value, and *A*1, *A*2, and *A*3 remained constant. For the vowel /ɪ/, the starting amplitude values for *A*1, *A*2, *A*3, and *A*4 were 60 dB as well, and the modifications to either *A*2 or *A*4 continuum were as for /i/. The 60-dB starting value for subsequent variation of either *A*2 or *A*4 for either vowel /i/ or /ɪ/ was assumed as the baseline 0-dB level for stepwise amplitude adjustments.

The synthetic tokens were analyzed acoustically using the same LPC measurement procedure as for vowels in natural speech. Both the constant values for the levels unmodified and the stepwise adjustments for either *A*2 or *A*4 are listed in Table 2. As can be seen, the nominal 2-dB decrements from the baseline 0-dB level were not always equal to 2 dB at the synthesis output. The smaller values for /i/ result from influence of other formants, which was not corrected for at present. Because *F*2, *F*3, and *F*4 are more closely spaced, the influence of *A*3/*A*4 on *A*2, and *A*2/*A*3 on *A*4 was greater for /i/ than for /ɪ/. Examples of the endpoints for all four continua can be heard in Mm1 through Mm4 which contain .wav files.

Mm1. *A*2 continuum for /i/: tokens i_*A*2_1 and i_*A*2_9

Mm2. *A*2 continuum for /ɪ/: tokens ih_*A*2_1 and ih_*A*2_9

 1529-7853/05/6(3)/119/7/$22.50

Table. 2. Amplitude measurements for the 4-formant synthetic vowels /i/ and /ɪ/. The output values (in dB) for each synthetic continuum (*A*2 or *A*4) vary in nine steps; constant values are shown for the first token of each series.

	Vowel /i/						Vowel /ɪ/				
Token	dB-step	*A*1	*A*2	*A*3	*A*4	Token	dB-step	*A*1	*A*2	*A*3	*A*4
i_*A*2_1	0	60.3	50.7	42.0	42.5	ih_*A*2_1	0	61.6	50.6	41.7	41.4
i_*A*2_2	−2		49.5			ih_*A*2_2	−2		49.0		
i_*A*2_3	−4		47.9			ih_*A*2_3	−4		47.6		
i_*A*2_4	−6		46.4			ih_*A*2_4	−6		45.6		
i_*A*2_5	−8		44.9			ih_*A*2_5	−8		43.5		
i_*A*2_6	−10		43.3			ih_*A*2_6	−10		41.4		
i_*A*2_7	−12		41.3			ih_*A*2_7	−12		39.0		
i_*A*2_8	−14		39.2			ih_*A*2_8	−14		36.6		
i_*A*2_9	−16		38.0			ih_*A*2_9	−16		35.2		
i_*A*4_1	0	60.3	50.7	42.0	42.5	ih_*A*4_1	0	61.6	50.6	41.7	41.4
i_*A*4_2	−2				40.1	ih_*A*4_2	−2				39.0
i_*A*4_3	−4				38.7	ih_*A*4_3	−4				37.1
i_*A*4_4	−6				38.0	ih_*A*4_4	−6				35.5
i_*A*4_5	−8				35.5	ih_*A*4_5	−8				33.3
i_*A*4_6	−10				33.9	ih_*A*4_6	−10				31.3
i_*A*4_7	−12				32.9	ih_*A*4_7	−12				29.2
i_*A*4_8	−14				32.5	ih_*A*4_8	−14				27.1
i_*A*4_9	−16				32.4	ih_*A*4_9	−16				25.2

Mm3. *A*4 continuum for /i/: tokens i_*A*4_1 and i_*A*4_9

Mm4. *A*4 continuum for /ɪ/: tokens ih_*A*4_1 and ih_*A*4_9

2.2 Listeners

The listeners were 11 native speakers of Midwestern American English (six men and five women). They were students at Ohio State University, had pure-tone thresholds of 20 dB or better at octave intervals from 250–8000 Hz, and were paid for their participation.

2.3 Procedures

The four sets of stimuli blocked by vowel (/i/ or /ɪ/) and formant (*F*2 or *F*4) were presented to a listener seated in a sound-attenuating booth. Five randomized blocks were prepared for each set. In each block, the stimuli were arranged in pairs so that nine tokens occurred in all possible pair combinations in both AB and BA order, producing 72 pairs per block. Each listener responded to 80 presentations of each token for each set (16 presentations×5 blocks). Stimuli were presented over Sennheiser HD-414 headphones to the right ear of the listener at 70 dB SPL via TDT System II, after low-pass filtering at 5 kHz. Listeners responded to the stimuli in a two-interval 2AFC task by selecting one of two response box buttons displayed on a computer monitor. After listening to each pair, they indicated whether the first token in a pair was a more or less natural-sounding instance of the vowel than the second token (within the limitation of synthetic speech).

2.4 Results

Shown in Fig. 1 are mean responses for /i/ and for /ɪ/. Percentage of classifications as more natural-sounding vowel is displayed as a function of variation in either *A*2 or *A*4. For both vowels, subjects were sensitive to changes in *A*2 [Fig. 1(a)]. Higher amplitude values for /i/

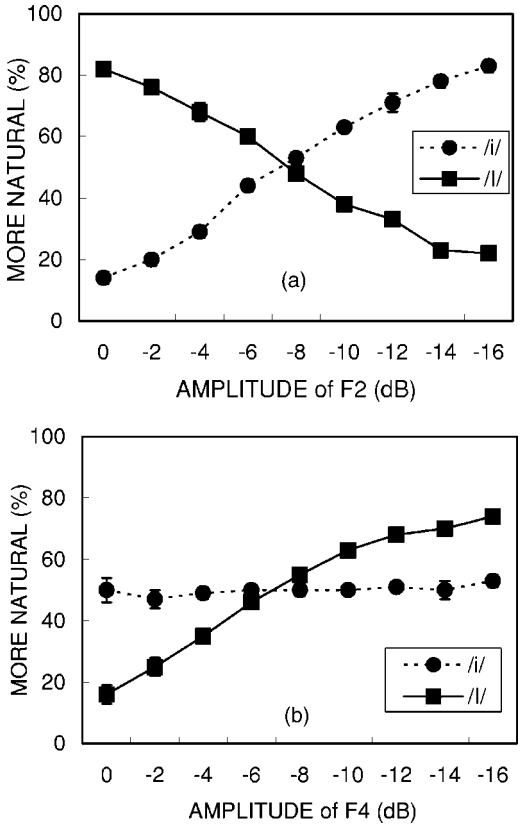

Fig. 1. Classification as more natural-sounding instances of the vowels /i/ and /ɪ/ as a function of changes in the relative amplitude of either *F*2 (a) or *F*4 (b) in experiment 1. Each data point represents average value for 11 listeners. Error bars show one standard error.

(tokens i_*A*2_1 through i_*A*2_4; see Table 2) yielded the lowest number of the natural-sounding /i/ which was perceived as more natural as *A*2 values decreased. The opposite was found for /ɪ/ in that the tokens with higher *A*2 values sounded more natural.

The responses to variation in *A*4 [Fig. 1(b)] indicate that, for /i/, the listeners were unable to differentiate between more- or less-natural sounding tokens, as apparent in the flat response function. However, they were sensitive to *A*4 variation for /ɪ/. The tokens with the highest amplitude values sounded less natural in comparison with the tokens at the lower end of the continuum.

A repeated-measures ANOVA with the factor formant level (*A*2 or *A*4) was performed for each vowel and each formant. The total number of "more natural-sounding /vowel/" responses was summed across all 9 steps for each subject. For /i/, there was a significant effect of *A*2 [$F(8,80)=141.1, p<0.001$] and no significant effect of *A*4 [$F(8,80)=0.418$, $p=0.907$]. For /ɪ/, both effects of *A*2 and *A*4 were significant [$F(8,80)=108.8, p<0.001$] and [$F(8,80)=81.5$, $p<0.001$, respectively]. These results show that variation in *A*2 for either vowel /i/ or /ɪ/ had an effect on perceived vowel naturalness. Similar changes to *A*4 had an effect on the naturalness of /ɪ/ but did not affect that of /i/.

Comparing listeners' responses with the amplitude variation in natural speech (NS), the following observations can be made. For the vowel /i/, listeners consistently perceived those tokens having *A*2 values closer to those in the NS as more natural-sounding instances of /i/. In

particular, the relative *A*2 (*A*2-*A*1) of the terminal token i_*A*2_9 was −22.3 dB, which approximated the mean Rel*A*2 listed in Table 1 (−22.4 dB). However, listener responses to the *A*4 series indicated a lack of sensitivity to changes in relative *A*4. The patterns of responses for /ɪ/ (for both *A*2 and *A*4 series) show a general agreement with variations in NS. The relative values of *A*2 (*A*2-*A*1) or *A*4 (*A*4-*A*1) closer to the endpoints of either continuum approximate Rel*A*2 and Rel*A*4 means in NS listed in Table 1. Overall, these results suggest that, except for *F*4 of /i/, listeners are sensitive to the variations in vowel formant amplitude in accord with a general trend observed in natural speech.

3. Experiment 2

Experiment 2 tested the effects of formant level variation on vowel identification as opposed to naturalness.

3.1 Procedures

The stimuli, listeners, and experimental setup were the same as in experiment 1. The only difference was that listeners responded to the stimuli in a single-interval 2AFC identification task with the choices "/i/" and "/ɪ/." Two randomized blocks per formant level were presented. In each block, each of the 9 tokens appeared 10 times for a total of 90 trials per block. The order of presentation was counterbalanced.

3.2 Results

Figure 2 displays mean identification functions for each formant level and each vowel. Listeners were able to identify each token as an instance of either /i/ or /ɪ/ as a function of a change in formant amplitude. For *A*2, the classification of /i/ was generally in line with its perceived naturalness: the less natural-sounding instances of /i/ were classified mostly as /ɪ/ and identifications as /i/ progressively increased with the decline in formant amplitude [Fig. 2(a)]. In both experiments, the best /ɪ/ tokens, in terms of either naturalness or identification, had the highest amplitudes of *F*2.

As in experiment 1, variation in *A*4 had no effect on listeners' identification of /i/ but did affect their judgments for /ɪ/ [Fig. 2(b)]. The vowel /ɪ/ tended to be perceived as an /i/ when *A*4 values were high and as an /ɪ/ when they were lower. A repeated-measures ANOVA with the factor formant level was performed for each vowel and each formant. The results were similar to the results of experiment 1. For /i/, the effect of *A*2 was significant [$F(8,80)=75.7, p<0.001$] but the effect of *A*4 was not [$F(8,80)=1.01$, $p=0.437$]. For /ɪ/, there was a significant effect of both *A*2 [$F(8,80)=34.7, p<0.001$] and *A*4 [$F(8,80)=27.02, p<0.001$].

4. Summary and conclusions

Listeners' sense of vowel naturalness seemed to be related to their vowel quality judgments. The more natural-sounding tokens yielded higher identifications rates and, conversely, the less natural-sounding ones were identified as instances of a different vowel. This implies that spectral details present in formant amplitude variations, presumably learned through language experience, affected some aspects of vowel identity.

The question arises as to the role this fine acoustic detail may play in vowel perception. Most importantly, can formant amplitude variations cue vowel quality distinctions in natural speech? The present results suggest only that listeners are sensitive to such variation under controlled laboratory conditions. Moreover, these experiments tested sensitivity to variation in the amplitude of one formant at a time, not considering a combined effect of such adjustments to more than one formant. In responding to synthetic vowels, listeners may focus on cues such as vowel brightness which may or may not be utilized in vowel perception across contexts and speakers. Whether and to what extent differences in vowel timbre contribute to vowel perception in communicative situations is an open question. More systematic experimentation is needed to explore and determine the role of this fine acoustic detail.

 1529-7853/05/6(3)/122/7/$22.50

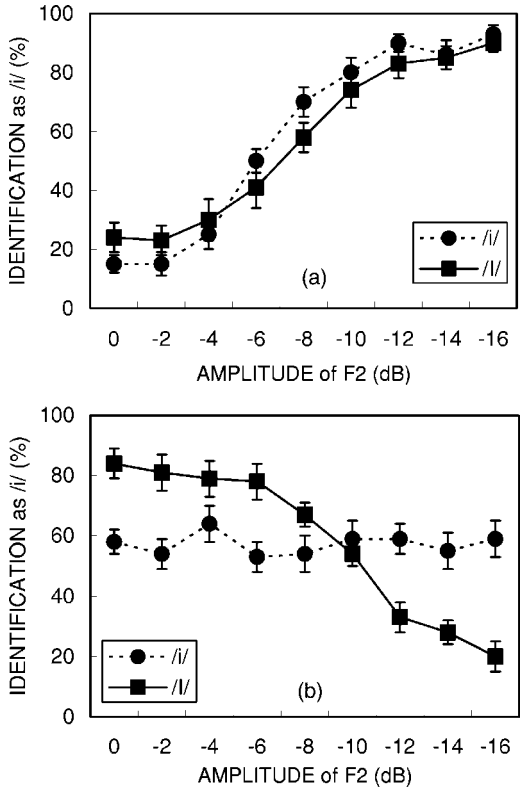

Fig. 2. Identification of the vowels /i/ and /ɪ/ as a function of changes in the relative amplitude of *F*2 (a) and *F*4 (b) in experiment 2. Identification of either vowel as /i/ is highest when the identification functions approach 100%. Identification of either vowel as /ɪ/ is highest when the identification functions are close to 0%, which corresponds to the lowest percentage of identification of either vowel as /i/. Each data point represents average value for 11 listeners. Error bars show one standard error.

Acknowledgments

This work was supported by NIH NRSA training grant (R. A. Fox., PI). I thank Robert A. Fox for his help and valuable discussions related to this project. I also thank James Hillenbrand and two anonymous reviewers for their insightful comments.

References and links

Aaltonen, O. 1985. "The effects of relative amplitude levels of *F*2 and *F*3 on the categorization of synthetic vowels," J. Phonetics **13**, 1–9.

Ainsworth, W. A., and Millar, J. 1972. "The effect of relative formant amplitude on the perceived identity of synthetic vowels," Lang Speech **15**, 328–341.

Carlson, R., Granström, B., and Fant, G. (1970). "Some studies concerning perception of isolated vowels," Speech Transmission Laboratory Quarterly Progress Status Report (STL-QPSR 2/3), 19–35.

Hawkins, S. 2003. "Roles and representations of systematic fine phonetic detail in speech understanding," J. Phonetics **31**, 373–405.

Hillenbrand, J. M., and Houde, R. A. 2003. "A narrow-band pattern-matching model of vowel perception," J. Acoust. Soc. Am. **113**, 1044–1055.

Ito, M., Tsuchida, J., and Yano, M. 2001. "On the effectiveness of whole spectral shape in vowel perception," J. Acoust. Soc. Am. **110**, 1141–1149.

Johnson, K., and Qi, Y. (**1987**). KLSYN: A formant synthesis program. Ohio State University, Columbus, OH.

Klatt, D. H. 1982. "Prediction of perceived phonetic distance from critical-band spectra: A first step," Proc. IEEE Int. Conf. Speech, Acoust. Signal Process., 1278–1281.

Lienard, J.-S., and Di Benedetto, M.-G. 1999. "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411–422.
Lindqvist, J., and Pauli, S. (1968). "The role of relative spectrum levels in vowel perception," Speech Transmission Laboratory Quarterly Progress Status Report (STL-QPSR 2/3), 12–15.
Miller, R. L. 1953. "Auditory tests with synthetic vowels," J. Acoust. Soc. Am. **25**, 114–121.
Peterson, G. E., and Lehiste, I. 1962. "Revised CNC lists for auditory test," J. Speech Hear Disord. **27**, 62–70.
TFR: Time Frequency Representation Software. (**1999**). Avaaz Innovations Inc.

