Estimating vowel formant discrimination thresholds using a single-interval classification task^{a)}

Eric Oglesbee^{b)} and Diane Kewley-Port

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

(Received 9 September 2008; revised 28 January 2009; accepted 29 January 2009)

Previous research estimating vowel formant discrimination thresholds in words and sentences has often employed a modified two-alternative-forced-choice (2AFC) task with adaptive tracking. Although this approach has produced stable data, the length and number of experimental sessions, as well as the unnaturalness of the task, limit generalizations of results to ordinary speech communication. In this exploratory study, a typical identification task was used to estimate vowel formant discrimination thresholds. Specifically, a signal detection theory approach was used to develop a method to estimate vowel formant discrimination thresholds from a quicker, more natural single-interval classification task. In experiment 1 “classification thresholds” for words in isolation and embedded in sentences were compared to previously collected 2AFC data. Experiment 2 used a within-subjects design to compare thresholds estimated from both classification and 2AFC tasks. Due to instabilities observed in the experiment 1 sentence data, experiment 2 examined only isolated words. Results from these experiments show that for isolated words, thresholds estimated using the classification procedure are comparable to those estimated using the 2AFC task. These results, as well as an analysis of several aspects of the classification procedure, support the viability of this new approach for estimating discrimination thresholds for speech stimuli.

© 2009 Acoustical Society of America. [DOI: 10.1121/1.3086269]

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

Pages: 2323–2335

I. INTRODUCTION

Previous work on examining vowel formant discrimination has often employed a modified two-alternative-forced-choice (2AFC) task with an adaptive tracking procedure (Kewley-Port and Watson, 1994; Kewley-Port *et al.*, 1996; Kewley-Port and Zheng, 1999; Kewley-Port, 2001; Liu and Kewley-Port, 2004b). Establishing thresholds for discriminating formants is important for determining the baseline ability of the auditory system to process small differences between vowels. Results of this research have shown how formant thresholds change when variables such as linguistic context, fundamental frequency, and hearing status were manipulated using a 2AFC task. While this modified 2AFC adaptive tracking task has been successful in estimating stable vowel formant thresholds, repeated stimulus presentations are quite unnatural compared to the single presentation of utterances that occurs in normal communication. Also, the number of trials necessary to obtain stable thresholds is rather large, and consequently listeners have to participate in multiple experimental sessions over the course of several days. Given these drawbacks, the purpose of the current study was to develop a method that used a linguistic labeling

task to estimate vowel formant thresholds comparable to those estimated using a modified 2AFC adaptive tracking paradigm.

Although it is a worthy goal to improve experimental procedures for measuring discrimination for any listener, a particular goal in our laboratory has been to understand vowel processing abilities first in young normal hearing (YNH) listeners and then to determine how age, hearing impairment, or first language might degrade those abilities. Because the unnatural 2AFC task might limit generalizing laboratory results to more ordinary communication by elderly hearing-impaired (EHI) or second language (L2) learners of English, we were motivated to develop more natural methods for obtaining thresholds. Reviewing some of our research to date, several studies have used 2AFC procedures to document how the ability to process isolated vowels degrades with typical high-frequency sloping hearing loss. In the first study, Coughlin *et al.* (1998) examined isolated steady-state vowels and found that vowel formant discrimination abilities for low-frequency formants (F1) were the same between EHI listeners and YNH listeners, but that thresholds were significantly elevated for high-frequency formants (F2). In an effort to relate these thresholds to the identification of vowels, moderate correlations were found between the F2 thresholds and vowel identification by EHI listeners, indicating that there is a relationship between reduced ability to correctly identify vowels and poorer discrimination ability.

In two follow-up studies also using isolated steady-state vowels, Richie *et al.* (2003) and Richie and Kewley-Port (2005) reported similar results for young hearing-impaired

^{a)} Portions of the data were presented at the 150th Meeting of the Acoustical Society of America (J. Acoust. Soc. Am. **118**, 1929–1930) and the Fourth Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan (J. Acoust. Soc. Am. **120**, 3129).

^{b)} Author to whom correspondence should be addressed. Electronic mail: eric.oglesbee@bethelcollege.edu. Present address: Bethel College, Department of Modern Language Studies, Mishawaka, IN 46545.

listeners. Given that such vowel stimuli are far different from those found in fluent sentences, new synthesis techniques using STRAIGHT (Kawahara *et al.*, 1999) were employed by Liu and Kewley-Port (2004b, 2007) to generate nearly natural vowel stimuli for three nine-word sentences. Their study of EHI listeners examined vowel formant discrimination for words in the sentences, as well as computer edited versions that presented only the word or only the isolated vowel from the word. Their results demonstrated a 100% elevation in F2 formant thresholds for EHI compared to YNH listeners for syllables and sentences. Encouraging results using the nearly natural speech stimuli from STRAIGHT led us to another set of experiments examining the relation between vowel thresholds and identification of vowels, this time with L2 learners of American English (AE). Listeners from three L2 languages with different vowel systems heard AE words. Preliminary results reported by Kewley-Port *et al.* (2005) showed little difference between the abilities of L2 and AE listeners to discriminate vowel formants. This is an important outcome, suggesting that more peripheral level processing abilities for vowels are not compromised by one's first language. However, outcomes of these L2 experiments, as well as those with EHI listeners, would be much stronger if the experimental tasks were more ecologically valid. The present study had the ambitious goal of designing more natural experimental procedures that could be used with many listener populations to determine the just noticeable differences in vowel spectra from a simple identification task [i.e., listen to a word (sentence) once and indicate what word was heard]. It should be noted that since the task we have chosen is an identification/classification approach for threshold estimation, it might not be appropriate for populations where large variability has made connecting identification performance to discrimination performance problematic (e.g., cochlear implant users; Iverson, 2003).

The task chosen for this study is a single-interval standard classification task using a continuum of stimuli that vary in one speech parameter, here the frequency of one vowel formant. Classification tasks (also called "labeling" or "identification" tasks) have long been used to examine the relation between category boundaries and discrimination in speech continua [e.g., voice onset time (VOT) continua from /ba/-/pa/, Lisker and Abramson, 1970]. Specifically, the prediction of discrimination performance from a classification task is at the core of research and theory concerning categorical perception (Harnad, 1987). However, there are clear differences between the goals of the current study and those related to categorical perception. For example, categorical perception is concerned with how discrimination performance changes along a classification continuum, especially at the boundary between categories (Repp, 1984), whereas the current study examines discrimination near a category center. Therefore, although the current study examines discrimination using a classification task, our interest is in determining the just noticeable difference between stimuli in contrast to the categorical perception literature that examines the nature of phonetic categories.

Some categorical perception studies have examined how details of the discrimination task alter the observed discrimination function and, therefore, the relationship between discrimination and predictions from the classification task. For example, if the stimulus properties are more steady state (vowel-like) versus rapidly changing (consonant-like), observed discrimination functions are more uniform and performance exceeds predicted discrimination (Pisoni, 1973; Mirman *et al.*, 2004). Also, changes in the discrimination trials from ABX to more sensitive tasks that use less working memory, such as the 4IAX (Pisoni, 1975) or 4I2AFC (Gerrits and Schouten, 2004) tasks, tend to flatten the discrimination function. While some of these issues are certainly related to the present experiments, these tasks all use two or more presentations of the stimuli per trial to measure discrimination. This contrasts with our goal of estimating discrimination thresholds using a single presentation of a stimulus by referencing the underlying psychometric function.

In the current study, principles from "Thurstonian scaling," as implemented in signal detection theory (SDT) (Macmillan and Creelman, 2005, p. 155), are used to find a point on a classification psychometric where thresholds comparable to those found in a 2AFC task can be estimated. Previously established theoretical relationships between performances on a 2AFC adaptive tracking task using a two-down-one-up (2D1U) rule (Levitt, 1971) and a two-category classification task (two-category classification is functionally equivalent to a yes/no decision) are used to provide a range of sensible cumulative d' values for estimating an empirical threshold from the classification psychometric. Thus, two types of experimental tasks were employed in the present study. In experiments 1 and 2 listeners participated in a single-interval classification task. In this task, a stimulus was presented, and listeners categorized the stimulus as one of two response categories (possible response sets were *bid/bed*, *cut/cot*, and *hack/hawk* in the experiments reported here). In experiment 2, the same listeners also used a modified 2AFC adaptive tracking task. This modified 2AFC adaptive tracking task had the same format as the one used in the previous discrimination studies cited earlier, in particular, Kewley-Port and Watson (1994). On each trial listeners heard three stimuli. The first stimulus presented on each trial was always the "standard." The second and third intervals contained the standard and test stimuli played in random order. Consequently, the possible stimulus sequences were <STS> and <SST>, where "S" was a fixed standard for each formant continuum and "T" was the test stimulus. Listeners were asked to decide if the second or third interval contained the test stimulus. Using the typical 2D1U adaptive tracking rule, stable performance was obtained after a large number of trials (here 504 per formant) to measure thresholds from mean reversals.

To generate more natural speech stimuli for the classification task, stimuli were synthesized by using STRAIGHT (Kawahara *et al.*, 1999). STRAIGHT is a powerful speech analysis routine implemented in MATLAB that allows experimenters to manipulate amplitude, fundamental frequency, temporal relationships, and spectral characteristics of natural speech tokens. Although many of our earlier vowel threshold

TABLE I. Acoustic characteristics of stimuli used in experiments 1 and 2.

	F1 (Hz)	F2 (Hz)	F3 (Hz)	Midpoint F0 (Hz)	Duration (ms)
bid	345	2196	2961	183	200
bed	485	1927	2918	162	206
cut	668	1561	2692	160	91
cot	829	1432	2616	157	171
hack	818	1906	2530	161	190
hawk	700	1055	2643	162	190

experiments used more controlled vowels (albeit unnatural sounding) created with the KLTSYN synthesizer (Klatt, 1980). Liu and Kewley-Port (2004a) showed that formant thresholds were very similar when estimated using STRAIGHT-synthesized as compared to formant-synthesized speech for analogous vowel stimuli.

In experiment 1, formant discrimination thresholds for vowels in both isolated words and in sentences were estimated using a classification task and then compared to previously collected 2AFC data (Liu and Kewley-Port, 2004b). As a follow-up, experiment 2 implemented a within-subjects design to estimate these thresholds using both a classification and a 2AFC task. Because unstable data were obtained from the sentence stimuli in experiment 1, only isolated words were examined in experiment 2. The isolated word results from the two experiments indicate that there are common processing mechanisms underlying vowel discrimination and classification that allow vowel discrimination performance to be predicted from classification performance. Furthermore, we present an analysis of the relationship between classification function parameters and estimated thresholds for a particular d' criterion, along with some suggested guidelines for implementing the proposed formant discrimination threshold estimation method.

II. EXPERIMENT 1

A. Method

1. Stimuli

The English vowels / I /, / E /, / A /, and / a / were chosen for investigation. These vowels were examined in several previous discrimination studies (Kewley-Port and Watson, 1994; Kewley-Port *et al.*, 1996; Kewley-Port and Zheng, 1999; Kewley-Port, 2001; Liu and Kewley-Port, 2004b) that used stimuli generated with either the Klatt (1980) synthesizer or STRAIGHT (Kawahara *et al.*, 1999). In this experiment, using STRAIGHT, four seven-step continua were resynthesized from words taken from recordings of a female speaker who read a list of nine-word sentences that had been designed to have relatively neutral prosody. The speaker for this study was the same one used as a model for synthesis in the studies listed above. The sentences were constructed such that the second, fifth, and eighth word positions within the sentence were filled with combinations of words taken from the minimal pairs *bid/bed*, *cut/cot*, and *hack/hawk*. Listening and acoustic analysis were used to select one prosodically neutral production for each word. Table I contains formant, F0, and vowel duration information for the six selected

FIG. 1. F1 and F2 values for the four continua used in the forced choice classification (or labeling) tasks in experiments 1 and 2. *Bid/bed* continua are shown in the top panel; *cut/cot* continua are shown in the bottom panel. F1 and F2 shifted continua are indicated by closed and open circles, respectively. Formant values for base stimuli are shown with a downward pointing triangle. Reference stimuli for shifted end points are shown with an upward pointing triangle. Note that for the F2 continua, F1 was first shifted to an intermediate position between the base and reference stimuli.

words. The words *bid* and *cut* were used as base stimuli for resynthesis, while *bed* and *cot* were used as reference stimuli for developing classification continua. *Hack* and *hawk* were used as fillers in the test sentences.

Figure 1 contains a graphical representation of the parameters modified in STRAIGHT across the four classification continua. In each continuum, as in the previous vowel formant discrimination studies cited earlier, a single formant was shifted in equal steps in hertz, while all other acoustic properties were held constant. The direction and magnitude of the formant shifts were determined by informal listening such that the base stimuli *bid* and *cut* would be well-identified as *bed* and *cot*, respectively, for the maximum shift. One strategy was to shift formants to more extreme values than observed in the reference stimuli. Also, to obtain high quality F2 stimuli, F1 was shifted to a somewhat ambiguous value before F2 was shifted. Because the original *bid* token had a relatively high F0, *bid/bed* tokens had F0 lowered by 10% in STRAIGHT's F0 resynthesis routine. To neutralize the duration difference between *cut* and *cot*, duration was lengthened by 20% for the base *cut* stimulus.

The procedure for shifting formant frequencies was a very similar version of the routine employed by Liu and Kewley-Port (2004b). First, a matrix in MATLAB representing the spectrogram (row=frequency; column=time; matrix entry=amplitude) of the base word was obtained by STRAIGHT (Kawahara *et al.*, 1999). Second, to shift a formant peak, the temporal onset and offset of the formant (including transitions) were visually identified. Third, in each time frame (i.e., one spectrum), the bottom of the valleys on either side of the formant peak was selected. Amplitude in the valley opposite of where the formant was shifted was adjusted to be a constant level that was equal to the bottom of the valley, while the shifted peak replaced the remaining spectral values. In most cases, the shift did not result in changes to other formants. A rule was employed where the shifted and original amplitude values were compared, and whichever was larger was retained. This resulted in formants being smoothly merged with one another when shifted to close proximity. Thus, details in the formant peaks were preserved in this procedure, with the valleys only somewhat changed. At last, this modified matrix was reloaded into STRAIGHT, and when necessary, other acoustic modifications such as F0 lowering and temporal expansion were simultaneously implemented [for more details on formant-peak shifting, see Liu and Kewley-Port (2004a) and their Fig. 1].

Sentence stimuli were created by digitally inserting words into the low-predictability sentence frame “Write _____ along with _____ but not _____ please,” which was the same frame that was used to elicit the words during recording. Depending on the continuum that was being tested (*bid/bed* or *cut/cot*), the other positions were filled with a word from the pair *hack/hawk* and either *bid/bed* or *cut/cot*. All non-shifted words in the sentence were processed by STRAIGHT with no parameter changes such that the acoustic quality introduced by the STRAIGHT resynthesis was similar for all words tested.

2. Listeners

Indiana University students with normal speech and hearing by self report were paid for their participation in the present study. Of the 35 students tested, data from three listeners (two *bid/bed* F2 and one *cut/cot* F2) were excluded because the listeners had near chance identification of the isolated word end-point stimuli.

3. Procedures

Stimuli were presented to the right ears of listeners, who were seated in a sound-treated IAC booth, via TDH-39 earphones. Stimulus presentation was controlled by TDT System II modules including a 16-bit digital/analog converter, a programmable filter, and a headphone buffer. A low-pass filter with a cutoff frequency of 4500 Hz, and an attenuation level set by the calibration procedure, was loaded into the programmable filter. The standard vowel /e/ was iterated for a duration of 3 s and used for calibration. The sound pressure level (SPL) for the test stimuli measured in a NBS-9A 6-c³ coupler by a Larson-Davis sound meter (model 2800) using

the linear setting was adjusted to be 70 dB SPL, with variation between stimuli from the same continuum less than 1.1 dB SPL.

Listeners were randomly assigned to four groups containing eight listeners each. Each group was exposed to stimuli from only one of the generated continua in both word and sentence conditions (i.e., either F1 shifted *bid/bed*, F1 shifted *cut/cot*, F2 shifted *bid/bed*, or F2 shifted *cut/cot*). Four types of blocks were run: word familiarization, word classification, sentence familiarization, and sentence classification. In each of these tasks, listeners were asked to indicate what word they heard. For isolated words, the possibilities were either *bid/bed* or *cut/cot*, depending on experimental group. For sentences, there were three words identified from the pairs *bid/bed*, *cut/cot*, and *hack/hawk* in each sentence. In the word familiarization blocks, listeners were exposed to the continuum end points in an alternating fashion for 12 trials with feedback for the correct response. In sentence familiarization, listeners heard the nine-word sentence, after which six response buttons appeared containing the words *hack*, *hawk*, *cut*, *cot*, *bid*, and *bed*. The order of presentation of the buttons matched the order of presentation of the word pairs in the sentence. Six sentences with continuum end points were used in the familiarization with feedback provided by the experimenter. The word and sentence classification blocks were similar to their familiarization counterparts, except that all seven stimuli from the continuum were presented multiple times in a random order, without feedback. In the word classification blocks, each stimulus was played nine times, resulting in 63 trials per block. The sentence classification blocks presented each stimulus twice in each sentence position, resulting in 42 trials per block.

Each subject participated in a single experimental session using a fixed protocol of block order. The number of blocks to approximate stable classification performance was determined from pilot experiments. Thus, altogether, four word familiarization, one sentence familiarization, four word classification, and six sentence classification blocks were presented. The session lasted for a total of 75–90 min, depending on a listener’s response latency. The last two word classification blocks and last three sentence classification blocks were used for analysis.

B. Results

1. Raw labeling functions

The results for each subject were calculated as the proportion of *bed* or *cot* responses obtained over the 18 presentations of each stimulus. Figures 2 and 3 display a representative set of raw labeling functions for four of the eight conditions. Figure 2 displays data from the *bid/bed* continuum for both F1 and F2 in the word condition. The reasonably consistent performance across listeners exhibited in Fig. 2 are representative of behaviors also observed for *bid/bed* (F1 and F2) in the sentence condition and *cut/cot* (F1 and F2) in the word condition. Listener performance for *cut/cot* (F1 and F2) in the sentence condition was much less consistent than the other conditions and is illustrated in Fig. 3. In general, raw labeling functions were more consistent

FIG. 2. Raw identification functions for *bid/bed* F1 and F2 stimuli in the word condition (18 judgments per stimulus). Each curve corresponds to a unique listener, numbered 1–16, with an additional identifier (F1 or F2) indicating which experimental group they belonged to. Each continuum contained seven stimuli that were randomized and presented to listeners for identification. The top panel contains results for F1 shifted stimuli; the bottom panel contains results for F2 stimuli. In both panels, stimuli 1 and 7 correspond to the most *bid*- and *bed*-like stimuli, respectively. Stimulus numbers correspond to their respective continua in the top panel of Fig. 1 (i.e., stimulus numbers in the top panel of Fig. 2 correspond to the F1 shift range of the top panel in Fig. 1). Below each panel, the shifted formant peak difference between stimulus steps is given in Hz.

and categorical for the word condition with F1 shifted formants compared to the sentence condition with F2 shifted formants.

2. Threshold calculation method

For discrimination, the concept of a “threshold” is the smallest amount of difference between two sounds that can be reliably detected. In most previous experiments reported by Kewley-Port and her colleagues (Kewley-Port and Watson, 1994; Liu and Kewley-Port, 2004b), vowel formant thresholds were measured as ΔF in hertz using a 2D1U rule in a modified 2AFC adaptive tracking paradigm. These 2AFC thresholds were calculated from average reversals along a stimulus continuum. Mathematically, this corresponds to the 70.7% correct point on a psychometric function plotted between 50% chance and 100% (Levitt, 1971).

FIG. 3. Raw identification functions for *cut/cot* F1 and F2 stimuli in the sentence condition (18 judgments per stimulus). Each curve corresponds to a unique listener, numbered 17–32, with an additional identifier (F1 or F2) indicating which experimental group they belonged to. Each continuum contained seven stimuli that were randomized and presented to listeners for identification. The top panel contains results for F1 shifted stimuli; the bottom panel contains results for F2 stimuli. In both panels, stimuli 1 and 7 correspond to the most *cut*- and *cot*-like stimuli, respectively. Stimulus numbers correspond to their respective continua in the bottom panel of Fig. 1 (i.e., stimulus numbers in the top panel of Fig. 3 correspond to the F1 shift range of the bottom panel in Fig. 1). Below each panel, the shifted formant peak difference between stimulus steps is given in Hz.

In this paper, the term “classification threshold” is used to indicate estimations of discrimination thresholds obtained from the classification task. A difference between thresholds obtained from 2AFC adaptive tracking and classification tasks is that the psychometric function ranges from 50%–100% in a 2AFC task as compared to 0%–100% in a classification task. The goal of the present research was to establish a method to calculate a classification threshold that corresponds well to the 70.7% threshold criterion in a 2AFC discrimination task. This was approached using a SDT implementation of Thurstonian scaling for one-dimensional classification tasks (Macmillan and Creelman, 2005, pp. 114–119). In this approach, a threshold criterion in terms of cumulative d' (measured from a classification continuum end point) is selected. Then, an empirical threshold is calculated by locating the point on the classification psychometric that corresponds to the chosen cumulative d' criterion. In prin-

ciple, the choice of the criterion used to calculate the empirical threshold is arbitrary, although the threshold criteria used in the current experiments were selected with the goal of obtaining results comparable to previously reported vowel formant discrimination thresholds for YNH listeners (Liu and Kewley-Port, 2004b).

The following method was used in the current experiments for deriving an estimation of a discrimination threshold from a two-category classification task (i.e., a classification threshold). It should be noted that this method assumes that the end points of the classification continua were well-identified by a majority of listeners. First, using a least-squares method, a logistic function [see Eq. (1)] was fit to the labeling data,

$$p(s) = 1/(1 + e^{((A-s)/K)}), \quad (1)$$

where A and K are model parameters representing the 50% point and slope of the function, respectively, and s is the stimulus number (in the current experiment s ranged from 1 to 7). Next, the value for $s_{\text{threshold}}$ resulting in $z(p(s)) - z(p(1)) = D$ is determined, where z is the inverse of the normal distribution function and D is the d' value chosen for estimating the empirical threshold (a discussion of the D used in the current experiments appears below). Using the known step size in hertz between stimuli and assuming that the range for s begins at 1, ΔF is calculated using

$$\Delta F = (s_{\text{threshold}} - 1) \times \text{step size} \quad (2)$$

This value, ΔF , is the classification threshold calculated separately for each listener in each condition.

Because the choice of the cumulative d' criterion (i.e., D) associated with a threshold in a one-dimensional classification task is arbitrary, choosing D depends on the goals of the experiment. In the current experiments, the goal was to have the classification thresholds be comparable to those measured in a modified 2AFC adaptive tracking task. Two options presented themselves. One possibility was to collect classification data and then in a *post hoc* analysis determine a value for D which would maximize the correspondence in performance between two tasks. Given the relatively small amount of previous data available, this approach was not pursued. Therefore, we chose a second option, which was to use previously established mathematical relationships between our types of tasks (Macmillan and Creelman, 2005) in an effort to derive a set of candidate values for D and to then determine which of these D -values produced the best overall correspondence between the two tasks. In particular, d' conversions between 2AFC, yes/no, and “reminder” discrimination tasks were selected.

Because a traditional 2AFC adaptive tracking paradigm using a 2D1U rule estimates thresholds at the 70.7% level on the psychometric function, the d' between the 50% chance and 70.7% threshold level was used as the “basic” choice for D [$D_{\text{basic}} = z(0.71) - z(0.5) = 0.553$]. However, SDT also derives a relationship between 2AFC and yes/no tasks (we note that our two-category classification task is functionally equivalent to a yes/no task; i.e., a *bid/bed* decision could be interpreted as either *bid* or *not bid*). In a 2AFC task, d' is $\sqrt{2}$ times that of a yes/no task (Macmillan and Creelman, 2005,

pp. 166–168). Thus a second possible cumulative d' criterion is $D_{Y/N}$, where $D_{Y/N} = D_{\text{basic}} \times 1/\sqrt{2} = 0.391$. A third possibility also presents itself. Experiment 2 used a modified 2AFC task, with the standard stimulus appearing in the first interval on every trial ($\langle\text{STS}\rangle$ and $\langle\text{SST}\rangle$ were the only possible stimulus sequences). One plausible strategy for listeners would be to ignore the third interval and simply attend to the stimuli in the first and second intervals, respectively, since these two intervals contain enough information to make the discrimination decision ($\langle\text{ST}\rangle$ =test stimulus in the second interval; $\langle\text{SS}\rangle$ =test stimulus in the third interval). This strategy is equivalent to interpreting the modified 2AFC task as a “reminder paradigm,” where if the listener uses a “differencing” strategy as defined by Macmillan and Creelman (2005), pp. 180–181, the d' for the 2AFC task would be $1/\sqrt{2}$ times that of a reminder task. Thus the third cumulative d' criterion is $D_{\text{remind}} = D_{\text{basic}} \times \sqrt{2} = 0.783$. Analysis of which of these D (D_{basic} , $D_{Y/N}$, and D_{remind}) result in the best fit to 2AFC thresholds is deferred to experiment 2 where a within-subjects comparison was possible. Given that the outcome of that analysis demonstrated that $D_{\text{remind}} = 0.783$ had the best match, classification thresholds in experiment 1 are reported using this cumulative d' criterion.

It is important to note that the possible relationships between the classification and modified 2AFC tasks were used only as a guide in selecting D . The d' relationship between 2AFC, Y/N, and reminder tasks discussed by Macmillan and Creelman (2005) is in the context of discrimination experiments using identical stimulus sets. More specifically, our experiments differ in two fundamental aspects:

- (1) One of our tasks was a two-category *classification* task.
- (2) The tasks being compared have different stimulus sets.

Because of these two major differences, it is not clear that previously established relationships between d' in these tasks should hold in our application. We argue, however, that because the current study is exploratory in nature, the above conversion factors suggest a reasonable range for possible values of D .

3. Thresholds for words and sentences

Logistic functions were fit to all raw data functions, and ΔF thresholds [see Eqs. (1) and (2)] were calculated for $D = 0.783$ in Excel using the method described above. Calculated threshold averages and standard deviations for each formant in both the word and sentence conditions are given in Table II. Given the small number of listeners, as well as the unequal variance for word and sentence conditions, a non-parametric sign test (two-tailed) was used to examine separately for each formant whether the word and sentence thresholds within subjects were significantly different. Differences were not found for *bid/bed* F1 ($p < 0.29$) or F2 ($p < 0.73$); however, a statistically significant difference was observed for *cut/cot* F1 ($p < 0.01$) and F2 ($p < 0.01$). For both *cut/cot* F1 and F2, every subject had a higher threshold in the sentence task as compared to the word task. As expected, thresholds for words in sentences were equal to or greater than those for words in isolation.

TABLE II. Formant frequencies, thresholds, and standard deviation of ΔF (Hz) estimated using a classification task for four formants in two phonetic contexts: words and sentences.

	/ɪ/-F1		/ʌ/-F1		/ʌ/-F2		/ɪ/-F2	
Formant frequency	349		648		1564		2196	
	<i>M</i>	s.d.	<i>M</i>	s.d.	<i>M</i>	s.d.	<i>M</i>	s.d.
Word	34.8	6.3	45.4	9.5	73.8	25.1	107.7	29.9
Sentence	38.7	20.9	98.8	68.8	193.7	84.0	137.6	71.8

C. Discussion of experiment 1

1. Comparison to previous research using 2AFC adaptive tracking procedures

One goal for experiment 1 was to develop a method for estimating formant thresholds using a classification task, such that the estimated thresholds would be equivalent to those calculated using a 2AFC paradigm. The previous report of Liu and Kewley-Port (2004b) who determined F1 and F2 formant thresholds based on a 2AFC paradigm for the vowels /ɪ/ and /ʌ/ in /bVd/ contexts, both in isolation and in sentences, was influential in the design of the present experiment. Figure 4 contains individual listener thresholds from the syllable condition of that experiment, as well as the results of the word condition from experiment 1 (additional data from experiment 2 will be discussed later). Thresholds are very similar between Liu and Kewley-Port (2004b) and experiment 1, even though in the current experiment a /kVt/ context was used instead of /bVd/ for testing /ʌ/. Given that the number of listeners per formant was relatively small for the two experiments ($n=4$ and $n=8$) and that the number of listeners was quite different between the experiments, a non-parametric two-tailed Mann–Whitney U -test was used to test for differences between the 2AFC and classification data. In the word condition, thresholds for /ɪ/-F1 were found to be significantly different ($U=0, p<0.01$): every threshold estimated from the classification task was lower than those ob-

served in the previous 2AFC data. No statistical differences were found for /ʌ/-F1 ($U=11, p=0.460$), /ʌ/-F2 ($U=13, p=0.682$), and /ɪ/-F2 ($U=15, p=0.934$).

Unfortunately, however, in the sentence condition, thresholds for /ɪ/-F1 and /ʌ/-F1 from the two studies were significantly different ($U=4, p<0.05$ for each comparison). While no statistical differences were observed for /ʌ/-F2 ($U=6, p=0.110$) and /ɪ/-F2 ($U=16, p=1$), the range of values obtained in the classification task for /ʌ/-F2 was substantially greater than what was observed in the 2AFC experiment. In general, thresholds obtained from the classification task in the word condition appeared to match the previous 2AFC data better than those estimated from the sentence condition.

2. Positional effects within sentences

Differences in end-point categorization performance and estimated thresholds (Table II) between the word and sentence conditions using *cut/cot* stimuli seemed to indicate that some property of the sentence frame influenced categorization responses and, consequently, estimated thresholds. This contrasts with Liu and Kewley-Port (2004b) where no significant difference in performance was found between the word and sentence conditions. However, in the current study, striking differences for some individual listeners were observed.

Based on the results of Liu and Kewley-Port (2004b) where differences in word position did not significantly affect formant thresholds, by design our identification functions in the sentence condition were originally pooled across the three positions in the sentence frame. However, given the apparent differences between word and sentence conditions in this experiment, the sentence data were further analyzed by sentence position. This analysis using data from the last five of the six sentence blocks revealed one extreme example of position effects (listener F1_22). When *cut/cot* stimuli were placed in the final sentence position, F1_22 exhibited a complete reversal in the identification function. Stimuli that were identified as *cut* in initial and medial sentence positions were labeled as *cot* when presented in the final sentence position, and vice-versa.

Although F1_22 represents the most dramatic case observed among all participants, a visual inspection of the raw position functions across all listeners suggests that stimulus location within sentences had at least a moderate effect on the identification functions of 19 of the other 31 participants. It has been shown that the phonetic content of immediately

FIG. 4. ΔF for F1 and F2 of the vowels /ɪ/ and /ʌ/ estimated in isolated words. Vowel formants are shown from left to right in order of increasing frequency. Thresholds from individual listeners were estimated using a classification task (experiments 1 and 2) and a 2AFC paradigm (Liu and Kewley-Port, 2004b and experiment 2).

FIG. 5. Contour plots of estimated thresholds as a function of fitted logistic parameters and step size in Hz between stimuli in a classification task using a seven-step continuum. The step at which the fitted logistic reaches the 50% level (A parameter) is plotted on the horizontal axis, and the slope of the fitted logistic (K parameter) is plotted on the vertical axis. Each panel contains contours calculated from (A, K) parameter pairs. Contour lines are labeled with the estimated thresholds based on stimulus step sizes in Hz (shown above each panel) used in experiments 1 and 2. The black boxes in the figure indicates target ranges for (A, K) pairs for obtaining well-behaved threshold functions. Observed (A, K) parameter pairs from experiments 1 (filled circles) and 2 (empty circles) are shown.

adjacent segments can shift vowel perception (Holt *et al.*, 2000). In our results we see a long-distance effect where segments in neighboring words appear to substantially influence categorization. We hypothesize that the likely source of interference is the naturally produced vowel /a/ in the word *not* that appears before the final test position in the frame sentence; however, pursuing this question is beyond the scope of this paper. Consequently, estimating discrimination thresholds using the sentence classification task we designed does not appear to be a viable approach at this time. However, it should be noted that these context effects are not applicable to the isolated word stimuli.

3. Properties of classification threshold estimation procedure

In this section, a preliminary analysis of the general properties of the classification threshold estimation procedure is presented, with design elements and results from the word condition of experiment 1 used to illustrate these gen-

eral properties. Basic guidelines for designing classification experiments that obtain reliable, stable thresholds are discussed, as well as the generalizability of the procedure. Finally, the success of experiment 1 is evaluated in light of this analysis.

a. General properties Based on the description of the threshold estimation method in Sec. II B 2 of experiment 1, there are three primary factors that affect values calculated for classification thresholds: parameters of the fitted identification function, cumulative d' criterion (D), and step size between stimuli. In the following analysis, the parameters of the logistic function [Eq. (1)] fit in experiment 1 are discussed. These parameters are easily interpretable; namely, A is the 50% point of the function and K is a measure of slope (values close to zero result in steep slopes). It should be noted that the threshold estimation method is general enough that other functions (e.g., Weibull) could be fit to the classification data.

Using the design parameters of experiment 1 (seven-step continua), Fig. 5 contains contour plots of estimated thresh-

olds for $D=D_{\text{remind}}=0.783$ as a function of logistic parameters (A parameter on horizontal axis and K parameter on vertical axis) and step size (different panels). Dashed contour lines at 15 and 30 Hz intervals for F1 and F2, respectively, show estimated thresholds. Filled circles mark the thresholds observed in the word condition of experiment 1.

Although these contour plots were created using a specific set experimental parameters, two general observations concerning the threshold estimation method can be made. First, in some regions of the contour plots there is a nonlinear relationship between logistic parameters and estimated thresholds. In particular, when the 50% point of the psychometric occurs early in the function ($A < 2$), or when the slope is extremely steep ($K < 0.1$) or shallow ($K > 2$), the relatively flat portions of the contours indicate that the location of the 50% point has little to no effect on the estimated threshold.

Second, for fixed values of A and K , increasing step size results in specific threshold contours being shifted downward in the $A \times K$ space. For example, the threshold contours in panels 1 (*bid/bed* F1) and 4 (*bid/bed* F2) in Fig. 5 were calculated using step sizes of 32.33 and 75.67 Hz, respectively. Comparing panel 1 to panel 4, the location of the 60 Hz contour line is shifted downward substantially. Note that although this apparently is a strong effect, the experimenter still has a wide range of choices for step size. SDT presumes there is a fixed underlying probability distribution for perceiving stimuli along a continuum. Thus, assuming listener responses will adjust to different step sizes, resulting in sets of logistic parameters that give comparable thresholds using our method. While research is needed in order to verify this hypothesis, the following example demonstrates that the hypothesis is reasonable. Suppose two continua testing the same formant were created from the same initial vowel with step sizes of 32.33 and 39.97 Hz respectively (these are the same step sizes for the *bid/bed* and *cut/cot* F1 panels in Fig. 5). All things being equal, if the same listener were tested on both continua with different step sizes, the slope of the psychometric function for the continuum with a step size of 39.97 Hz should be steeper than the continuum with a step size of 32.33 Hz. Also, because the continua had the same initial vowel, the 50% level of the continuum with the 39.97 Hz steps should be shifted leftward. Thus, the resulting thresholds should be nearly the same even when the physical step size differs.

b. Guidelines for experimental design The threshold contours in Fig. 5 have implications for experimental designs using the new threshold estimation method. Although the contour plots span a large range of A and K parameters, it appears that the range of parameters that should be considered acceptable is much smaller due to the nonlinearities in the threshold contours. In the case of a seven-step stimulus continuum as was used in experiment 1, the black box ($2 \leq A \leq 6$; $0.2 \leq K \leq 1$) in Fig. 5 represents the set of (A, K) parameter pairs that result in well-behaved threshold contours. This region also corresponds to the set of (A, K) parameter pairs that one would expect from a well-designed two-category identification task. Thus, the black box in Fig. 5 can be used as a *post hoc* evaluation for the appropriate-

ness of classification thresholds estimated from a seven-step continuum.

c. Generalizability Although the above analysis and plots in Fig. 5 illustrate results from a seven-step continuum, preliminary examination of the threshold contour space suggest that the threshold estimation method may be applicable to continua containing 5–20 stimuli. In these cases, the range of A parameters contained within the black box would need to be adjusted; however, the K range should remain fixed. Also, the nature of the classification task suggests that the proposed estimation method may be used for estimating thresholds from any type of stimulus continuum that is designed to span two distinct categories, not just speech. For example, in vision research, thresholds for color categories could be examined. Obviously, further research is needed to examine either the effect of different numbers of steps or the application of the method to other stimuli.

One inherent limitation imposed by using any classification task is that only primary acoustic cues for a contrast can be studied in isolation (e.g., F1 or a complex cue such as VOT). This requirement arises out of the fact that different levels of the cue chosen for study must be capable of causing listener's judgments of category membership to shift reliably. Consequently, this method would not be appropriate for estimating F3 or F4 discrimination thresholds. However, the method would still be applicable if one wanted to pursue "multidimensional discrimination thresholds" where multiple primary and secondary cues are covaried in a single classification continuum, as is typical with complex cues such as VOT. Thus a future study of thresholds for vowels where F1 and F2 covary is in line with our goal of developing a more ecologically valid approach to formant discrimination.

d. Relation to traditional classification measures and predictions of discrimination from classification With respect to classification tasks, the 50% point on the psychometric function is typically the focus of study; however, the proposed threshold estimation method instead focuses on the end points of the psychometric. In order to test whether or not the thresholds estimated in experiment 1 are substantially different from the 50% point, a correlational analysis between estimated thresholds and the location of the 50% point was performed. First, a bark scale transform was applied to the data so that all four formants could be included in the same analysis (Kewley-Port and Zheng, 1998). A very small correlation was found ($r = -0.107$), indicating that classification thresholds are not directly related to the more traditional 50% boundary.

With respect to predicting discrimination performance from a classification task, the slopes of psychometric functions have previously been related to peaks in ABX discrimination functions (Repp, 1984). We note that the current approach incorporates more information in the determination of a discrimination threshold than previous approaches, because it takes into account both the slope of the psychometric and the location of the 50% crossover point. Below, in experiment 2, it is demonstrated that these two traditional measures of boundary performance (psychometric slope and 50% crossover point) can be successfully used to estimate within-category (i.e., non-boundary) discrimination thresholds.

TABLE III. Shifted formant step sizes for 2AFC stimuli in experiment 2.

	/t/-F1	/ʌ/-F1	/ʌ/-F2	/t/-F2
Base formant frequency (Hz)	349	648	1564	2196
Step size (Hz)	2.97	5.51	12.51	18.89
% of base frequency	0.85%	0.85%	0.80%	0.86%

e. Evaluation of experiment 1 In the four panels of Fig. 5 the (A, K) parameter pairs obtained for all listeners from the word condition of experiment 1 are plotted with filled circles. For three of the listener groups (*bid/bed* F1, *cut/cot* F1, and *cut/cot* F2), all of the data points fall within the black box discussed above; however, three of the *bid/bed* F2 listeners fall outside the black box, suggesting that the *bid/bed* F2 stimuli were not as well-designed as the others. Overall though, even with the *bid/bed* F2 difficulties, experiment 1 appears to have been well-designed with respect to the properties of the threshold estimation method.

III. EXPERIMENT 2

Given that classification thresholds for words (but not sentences) appear reliable, the purpose of experiment 2 was to directly compare, using a within-subjects design, formant thresholds estimated from a classification task with those calculated using a modified 2AFC adaptive tracking procedure. The motivation for using a within-subjects design was threefold. First, the within-subjects design permitted an evaluation of the three cumulative d' criteria discussed earlier (D_{basic} , $D_{Y/N}$, and D_{remind}). Second, lingering questions remained concerning the comparability of thresholds estimated using the two methods given the numerous differences in experimental variables between Liu and Kewley-Port (2004b) and experiment 1 (i.e., number of subjects, base stimuli for resynthesis, etc.). Third, because of the large variability observed in experiment 1, we wanted to examine the degree to which performance on the two tasks corresponded within individual listeners.

A. Method

1. Stimuli

For the classification portion of experiment 2, the continua developed in experiment 1 were used, allowing for replication of the results from experiment 1. The continua used in experiment 1 needed to be changed for use in a 2AFC adaptive tracking task because the step size between stimuli was too large. Therefore, four modified formant shifted continua were created for the 2AFC task using the same formant shifting methods employed in experiment 1. Similar to Liu and Kewley-Port (2004b), each continuum contained 15 stimuli (including the standard), and step sizes between stimuli were kept between 0.7% and 1% of the formant frequency of the standard stimulus. For each continuum, the standard was identical to the initial point of the corresponding continuum in the classification task. The specific step sizes for each 2AFC continuum were chosen such that some of the points from the classification continuum would be contained in the 2AFC continuum. Table III contains the for-

mant frequencies for each standard, as well as the step sizes for each continuum.

2. Listeners

Eleven Indiana University students with normal speech and hearing by self report were paid for their participation in the present study. Three listeners did not finish the study due to scheduling difficulties and personal choice. Consequently, their data were excluded.

3. Procedure

Each listener participated in five experimental sessions that were completed within three weeks of the initial session. The first two sessions contained classification blocks, and the last three sessions contained 2AFC blocks. Calibration procedures were the same for experiment 2 as experiment 1.

a. Classification task In experiment 1, listeners were exposed to only one of the four stimulus continua. However, in experiment 2, each listener made classification judgments on all four word continua. For each continuum, two word familiarization and four word classification blocks described in experiment 1 were presented in a fixed order. Two continua were assigned in a quasi-random way for each experimental session. Two blocks with the same target words could not occur on the same test day (i.e., *bid/bed* F1 and F2 could not be presented on the same day). Also, each day had to contain one set of F1 blocks and one set of F2 blocks. With these restrictions, a unique continuum ordering was possible for each of the eight listeners.

Each experimental session during the classification portion of the experiment lasted for 0.5–0.75 h (depending on a listener's response latency). Data for estimating thresholds were typically taken from the last two blocks; however, in cases where one of the last two blocks was substantially different than the other, an earlier similar block was used for analysis.

b. 2AFC task The trials for the four vowel formants (F1 and F2 for both /t/ and /ʌ/) were randomly presented within each block (i.e., under medium stimulus uncertainty). The modified 2AFC paradigm of Kewley-Port and Watson (1994) that was described in Sec. I was employed. Each experimental session lasted for 0.75–1 h with 96 trials in each block. Seven blocks per session were administered for three experimental sessions ($N=21$ blocks). ΔF for each listener was averaged from the mean reversals over the last four blocks in which performance was judged as stable by visual inspection.

B. Results

1. Effect of different cumulative d' criteria

Figure 6 provides an illustration of the effect of the three cumulative d' criteria ($D_{\text{basic}}=0.553$, $D_{Y/N}=0.391$, and $D_{\text{remind}}=0.783$) discussed earlier on average classification thresholds. In order to quantify the best match for these three values of D , the 2AFC results were used as the expected value, and sums of squares of the differences were calculated across the four formants. For the F1 stimuli, D_{basic} was the best fit, while D_{remind} was the closest match for the F2 stimuli. Summing the squared differences across all four formants, D_{remind} provided the best overall match and therefore

FIG. 6. Average thresholds for /i/-F1 and /i/-F2, and /i/-F1 and /i/-F2, estimated using classification and 2AFC tasks (formant frequency in Hz shown on horizontal axis). Error bars indicating one standard deviation are shown for the 2AFC thresholds, as well as the thresholds estimated using a d' criterion of 0.783 in the classification task. Mean thresholds for d' criteria of 0.391 and 0.553 are shown without error bars.

was used for reporting classification thresholds in experiments 1 and 2. Although this analysis is limited in scope, note that the values of D selected for the current study appear to bracket a reasonable range of values.

2. Classification and 2AFC tasks

In the classification task, raw labeling functions were similar to those observed in experiment 1. Individual thresholds for ΔF were estimated from logistic functions, as described in experiment 1, and are displayed in Fig. 4 (squares) and Table IV.

The average thresholds for the classification and 2AFC tasks were remarkably similar (Fig. 6 and Table IV). Using a repeated measures analysis of variance, a significant main effect for formant was found [$F(3, 18)=55.232, p<0.001$], but not for task [$F(1, 6)=0.838, p=0.395$] or for the task $\times$ formant interaction [$F(3, 18)=0.290, p=0.832$]. This significant effect of formant is consistent with previous observations that ΔF increases as formant frequency increases (Kewley-Port and Watson, 1994). The non-significant effect of task indicated that the null hypothesis of no difference between the tasks could not be rejected.

3. Threshold estimates within individual listeners

The similarity of the group results does not indicate the relation between classification and 2AFC thresholds within individual listeners. To determine whether the quicker, more

natural classification task taps into the same underlying discrimination process as a 2AFC task, several analyses were performed.

First, a Spearman rank order correlation coefficient was used to determine if performance for listeners was ordered from best to worst similarly in the two tasks. Data for /i/-F1 and /i/-F1 conditions were excluded because variability in thresholds for the classification task appeared too small ($s.d. \leq 5.5$ Hz, Table IV, Fig. 4) for correlations to be meaningful. Correlation coefficients for both /i/-F2 ($N=8, \rho=-0.0238$) and /i/-F2 ($N=7, \rho=0.2143$) indicate that the order of listener performance was not similar across the two tasks. The second way of comparing performance on the two tasks used a two-tailed correlated t -test to evaluate the similarities in individual performance in the F2 conditions. The differences in threshold estimates were not significant for either /i/-F2 ($p=0.717$) or /i/-F2 ($p=0.621$). Taken together, thresholds for individual listeners are generally similar between the two tasks but are not close enough to preserve order from best to worst performance.

Additional analyses of data from all four formants were made after a bark transform removed the effect of the increase in formant frequency on thresholds (Kewley-Port and Zheng, 1998). The bark transformed signed difference between 2AFC and classification thresholds was chosen as the dependent variable for the analyses in order to preserve whether classification thresholds either under- or over-estimated the standard 2AFC thresholds, where zero indicates no difference. The average difference between the tasks was -0.063 barks ($s.d.=0.163$), indicating that individual thresholds were on average slightly lower for the 2AFC than the classification task as hypothesized. While -0.063 barks is small in relation to the 0.10 bark threshold for vowels tested under optimal listening conditions (Kewley-Port and Zheng, 1999; Liu and Kewley-Port, 2004b), the standard deviation of 0.163 is somewhat high.

Overall, there are several ways in which the 2AFC and classification thresholds were comparable. The group thresholds were very similar, with a tendency for classification thresholds to be higher. Although some listeners performed similarly on the two tasks, generally, individual listener thresholds were not ordered similarly from best to worst in the two tasks. Apparently listeners in the two procedures were influenced by separate factors that can result in large differences in individual performance. Thus while classification thresholds cannot be considered a substitute for 2AFC thresholds, it is reasonable to suggest that they would be useful as a performance measure in studies of individual dif-

TABLE IV. Means and standard deviations for classification and 2AFC thresholds estimated in experiment 2.

	/i/-F1		/i/-F1		/i/-F2		/i/-F2	
Formant frequency	349		648		1564		2196	
	<i>M</i>	s.d.	<i>M</i>	s.d.	<i>M</i>	s.d.	<i>M</i>	s.d.
Classification	35.4	4.3	44.9	5.5	75.7	24.6	160.6	74.9
2AFC	19.1	6.8	35.5	18.1	80.2	17.9	140.9	31.4

ferences, because for a given individual, classification thresholds were shown to be reliable in their own right (see next section).

C. Discussion of experiment 2

1. Reliability of word classification thresholds in experiments 1 and 2

Although some of the details of the experiment 2 protocol were different than experiment 1, the word stimuli that were used for the classification portion were identical to those used in the word condition of experiment 1. Therefore, the reliability of thresholds shown in Fig. 4 can be examined between listener groups. Using two-tailed *t*-tests assuming equal variance ($N=8$), no statistically significant differences were found between experiments 1 and 2 for any of the formants: *bid/bed* F1 ($p=0.811$), *cut/cot* F1 ($p=0.813$), *cut/cot* F2 ($p=0.877$), and *bid/bed* F2 ($p=0.088$). We note that variance was high and thresholds were not as stable for /i/-F2. Overall, classification thresholds for words appear reliable across the two different experimental groups.

2. Comparison of 2AFC thresholds with previous research

Given the difference in number of listeners per vowel formant between experiment 2 ($N=8$) and Liu and Kewley-Port (2004b) ($N=4$), a non-parametric two-tailed Mann-Whitney *U*-test was used to compare 2AFC thresholds obtained across experiments. No significant difference was found for /ʌ/-F1 ($U=15, p=0.934$), /ʌ/-F2 ($U=12, p=0.570$), and /i/-F2 ($U=8, p=0.214$); however, a significant difference was found for /i/-F1 ($U=0, p=0.004$). In the case of /i/-F1, the thresholds estimated in experiment 2 ($M=19.1$ Hz, $s.d.=6.8$) were less than half the size of those observed in Liu and Kewley-Port (2004b) ($M=49.2$ Hz, $s.d.=5.9$) (see Fig. 4). A likely source of the threshold difference was that even though the same speaker was used for stimulus resynthesis, different recordings were used as base stimuli in the two experiments. Acoustic analysis showed that the base tokens for /i/ were comparable between the two experiments in all ways except one: the difference in intensity between F1 and F2 (18.2 dB) for the experiment 2 /i/ base stimulus was greater than the difference between F1 and F2 (13.9 dB) in the /i/ base stimulus used in Liu and Kewley-Port (2004b). It is hypothesized that the greater prominence of F1 relative to F2 in experiment 2 made F1 of /i/ in the current experiment more salient, and therefore thresholds were lower. Also, the lower relative intensity of F2 may have contributed to the difficulty of developing well-identified /i/-F2 classification stimuli.

IV. SUMMARY AND CONCLUSIONS

A novel method for estimating discrimination thresholds from a classification task that are comparable to those obtained using a modified 2AFC adaptive tracking task was proposed using principles of Thurstonian scaling as implemented in SDT. In experiment 1, this method was tested using words both in isolation and in sentence context. F1 and

F2 thresholds for two vowel continua (/i/ to /ε/ and /ʌ/ to /a/) in a classification task were examined, and results showed the following:

- For isolated words, classification thresholds for three of the four formants tested were statistically similar to those from an earlier 2AFC task (Liu and Kewley-Port, 2004b).
- In the sentence context, only two of the four classification thresholds were statistically similar to thresholds reported from a 2AFC task (Liu and Kewley-Port, 2004b).
- The sentence context exhibited strong effects on vowel classification, especially for the /ʌ/-/a/ continua.

The implication of this final point is that while the goal of developing more natural methods for estimating discrimination performance in isolated words was met, an extension of these methods to sentences was not successful. However, in light of the promising results for classification thresholds for words, an analysis of the properties of the threshold estimation method was presented, and general guidelines for the use of this method were proposed. The simpler, more natural classification task, together with reasonably natural word length stimuli, should provide ecologically valid methods for examining vowel perception in general. It may be particularly useful for use with EHI and L2 listeners, although this should be experimentally established first.

In experiment 2, the goal was to further explore the relationship between thresholds estimated from classification and 2AFC adaptive tracking tasks by using a within-subjects design. Given the strong effect the sentence context had in experiment 1, the scope of experiment 2 was restricted to vowel classification in isolated words. The same stimuli from experiment 1 were used for the classification task in experiment 2, and continua for the 2AFC task had additional stimuli resulting in smaller steps between the stimuli used in experiment 1. Results from experiment 2 showed the following:

- The best match between thresholds estimated using the two tasks was found using a cumulative d' criterion (D) of 0.783, which corresponded to interpreting the modified 2AFC task as a “reminder design” in which listeners employed a differencing strategy.
- Thresholds were reliable across different groups of listeners; i.e., thresholds estimated in experiment 2 were comparable to those observed in experiment 1.
- At the group level, no statistical differences were observed between thresholds estimated using the two tasks.
- At the individual level, thresholds estimated using the two tasks were not significantly different; however, rank order correlation results suggested that individual differences between listeners influenced performance on the two tasks.
- Differences between thresholds reported in the current study and those reported in a previous study are probably related to differences in the acoustic properties of the base stimuli used to generate the continua in each study.

In summary, this study described a quicker, simpler threshold estimation method that appears to provide a reasonable approximation of thresholds estimated using a 2AFC

adaptive tracking task. Given that the proposed method is much more natural for speech stimuli and that the amount of listening time required to obtain stable thresholds is substantially less than a 2AFC task, estimating discrimination thresholds from a classification task appears to be a useful option.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health Grant No. DC-02229 to Indiana University. Special thanks are due to Morgan Dotts for assisting with the collection and handling of data in the second experiment.

- Coughlin, M., Kewley-Port, D., and Humes, L. (1998). "The relation between identification and discrimination of vowels in young and elderly listeners," *J. Acoust. Soc. Am.* **104**, 3597–3607.
- Gerrits, E., and Schouten, M. E. H. (2004). "Categorical perception depends on the discrimination task," *Percept. Psychophys.* **66**, 363–376.
- Harnad S., ed. (1987). *Categorical Perception: The Groundwork of Cognition* (Cambridge University Press, New York).
- Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). "Neighboring spectral content influences vowel identification," *J. Acoust. Soc. Am.* **108**, 710–722.
- Iverson, P. (2003). "Evaluating the function of phonetic perceptual phenomena within speech recognition: An examination of the perception of /d/—/t/ by adult cochlear implant users," *J. Acoust. Soc. Am.* **113**, 1056–1064.
- Kawahara, H., Masuda-Kastuse, I., and Cheveigne, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.* **27**, 187–207.
- Kewley-Port, D. (2001). "Vowel formant discrimination: Effects of stimulus uncertainty, consonantal context and training," *J. Acoust. Soc. Am.* **110**, 2141–2155.
- Kewley-Port, D., Bohn, O.-S., and Nishi, K. (2005). "The influence of different native language systems of vowel discrimination and identification," *J. Acoust. Soc. Am.* **117**, 2399.
- Kewley-Port, D., Li, X., Zheng, Y., and Neel, A. (1996). "Fundamental frequency effects on thresholds for vowel formant discrimination," *J. Acoust. Soc. Am.* **100**, 2462–2470.
- 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). "Auditory models of formant frequency discrimination for isolated vowels," *J. Acoust. Soc. Am.* **103**, 1654–1666.
- Kewley-Port, D., and Zheng, Y. (1999). "Vowel formant discrimination: Towards more ordinary listening conditions," *J. Acoust. Soc. Am.* **106**, 2945–2958.
- Klatt, D. H. (1980). "Software for cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Levitt, H. (1971). "Transformed up-down method in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- Lisker, L., and Abramson, A. (1970). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague, 1967, pp. 563–567.
- Liu, C., and Kewley-Port, D. (2004a). "STRAIGHT: A new speech resynthesizer for vowel formant discrimination," *ARLO* **5**, 31–36.
- Liu, C., and Kewley-Port, D. (2004b). "Vowel formant discrimination for high-fidelity speech," *J. Acoust. Soc. Am.* **116**, 1224–1233.
- Liu, C., and Kewley-Port, D. (2007). "Factors affecting vowel formant discrimination by hearing-impaired listeners," *J. Acoust. Soc. Am.* **122**, 2855–2864.
- Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide* (Lawrence Erlbaum Associates, NJ).
- Mirman, D., Holt, L. L., and McClelland, J. L. (2004). "Categorization and discrimination of non-speech sounds: Differences between steady-state and rapidly-changing acoustic cues," *J. Acoust. Soc. Am.* **116**, 1198–1207.
- Pisoni, D. B. (1973). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," *Percept. Psychophys.* **13**, 253–260.
- Pisoni, D. B. (1975). "Auditory short-term memory and vowel perception," *Mem. Cognit.* **3**, 7–18.
- Repp, B. H. (1984). "Categorical perception: Issues, methods, findings," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), Vol. **10**, pp. 243–335.
- Richie, C., and Kewley-Port, D. (2005). "Vowel perception by noise masked normal-hearing young adults," *J. Acoust. Soc. Am.* **118**, 1101–1110.
- Richie, C., Kewley-Port, D., and Coughlin, M. (2003). "Discrimination and identification of vowels by young, hearing-impaired adults," *J. Acoust. Soc. Am.* **114**, 2923–2933.