Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English

Alexander L. Francis[a)]
Department of Speech, Language and Hearing Sciences and Program in Linguistics, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47906

Natalya Kaganovich[b)]
Program in Linguistics, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47906

Courtney Driscoll-Huber
Department of Speech, Language and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47906

(Received 16 May 2007; revised 17 March 2008; accepted 21 May 2008)

In English, voiced and voiceless syllable-initial stop consonants differ in both fundamental frequency at the onset of voicing (onset F0) and voice onset time (VOT). Although both correlates, alone, can cue the voicing contrast, listeners weight VOT more heavily when both are available. Such differential weighting may arise from differences in the perceptual distance between voicing categories along the VOT versus onset F0 dimensions, or it may arise from a bias to pay more attention to VOT than to onset F0. The present experiment examines listeners' use of these two cues when classifying stimuli in which perceptual distance was artificially equated along the two dimensions. Listeners were also trained to categorize stimuli based on one cue at the expense of another. Equating perceptual distance eliminated the expected bias toward VOT before training, but successfully learning to base decisions more on VOT and less on onset F0 was easier than vice versa. Perceptual distance along both dimensions increased for both groups after training, but only VOT-trained listeners showed a decrease in Garner interference. Results lend qualified support to an attentional model of phonetic learning in which learning involves strategic redeployment of selective attention across integral acoustic cues. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2945161]

PACS number(s): 43.71.An, 43.71.Es, 43.71.Rt [PEI] Pages: 1234–1251

I. INTRODUCTION

The acoustic patterns of speech sounds are highly multidimensional, in the sense that multiple acoustic properties typically correlate with the production of a particular phonetic category. Most, if not all, of these correlates have the potential to function as perceptual cues to categorization under appropriate circumstances, but not all cues are weighted equally in a given contrast. There are at least two major reasons that listeners might prefer to make a particular phonetic judgment on the basis of one cue over another. On the one hand, the perceived difference between two phonetic categories might be greater along one contrastive dimension than the other. Alternatively, some cues may be privileged (for particular phonetic decisions) because of learned or innate biases in the way they are processed.

The multiplicity of cues to phonetic contrasts is well documented. For example, Lisker (1986) describes a wide variety of acoustic correlates that differ systematically between productions of intervocalic /p/ and /b/ in English. Most or all of these correlates have been shown to be sufficient to cue the perception of this contrast in syllable-initial position, even in the absence of other cues (Lisker, 1978), but we will focus on four that have been more intensively studied: Voice onset time (VOT; Abramson and Lisker, 1970), the fundamental frequency at the onset of voicing (onset F0; Haggard *et al.*, 1970; Haggard *et al.* 1981), the degree of delay in the onset of the first formant (F1 cutback or voiced transition duration; Stevens and Klatt, 1974) and the relative amplitude of any aspiration noise in the period between the burst release and the onset of voicing (Repp, 1979). Despite the multiplicity of sufficient cues to the English stop-consonant voicing contrast, when more than one of these cues are presented to listeners, a pattern of dominance appears that suggests that some correlates are better able to serve as cues (often called *primary cues*) than others (*secondary cues*), at least in specific phonetic contexts. For the purposes of this study, the most relevant observation is that VOT appears to dominate other cues to voicing of syllable-initial stop consonants in English (Raphael, 2005). In particular, a variety of studies have shown that, in this context, VOT is preferred over onset F0 (Abramson and Lisker, 1985; Gordon *et al.*, 1993; Lisker, 1978; Whalen *et al.* 1993; see Francis and Nusbaum, 2002 for discussion). However, although such patterns of relative dominance are generally agreed upon, there is little consensus regarding the

[a)]Author to whom correspondence should be addressed. Tel.: (765) 494-3815. Electronic mail: francisa@purdue.edu
[b)]Also at: Department of Speech, Language and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47906.

psychological basis for such apparent prioritization of one acoustic cue over another.

One factor of note in this regard is that the results of group studies on this topic (including the present one) may obscure the presence of real individual differences in the relative weighting of these two cues. For example, Haggard *et al.* (1970) found that onset F0 "can be of some importance, but the wide differences in performance between subjects show that it is unimportant for some listeners" (p. 616). Similarly, Massaro and Cohen (1976, 1977) found a range of individual differences in reliance on onset F0 as compared to VOT and fricative duration in a series of studies on the perception of voicing in syllable-initial fricatives. Such differences in individual listeners' weighting of normally covarying acoustic cues are consistent with other studies showing similar differences even in the perception of nonspeech cues (e.g., Lutfi and Liu, 2007), and clearly invite further study. However, the observation of individual differences in weighting still does not address the question of what might motivate the prioritization of one cue over another and to what degree such weighting might be changed by experience.

A. Perceptual weighting

One possible reason for the relative dominance of one cue over another is that the perceptual distance between two categories may be different along two different dimensions of contrast. For example, the perceptual distance between two prototypical exemplars of English /b/ and /p/ is quite large according to VOT and may be somewhat smaller according to onset F0.[1] In this case, listeners would be expected to give more weight to VOT than to onset F0, if only because the VOT differences are more easily distinguished. On the other hand, it is also possible that one dimension might be intrinsically better at attracting listeners' attention to it than another, such that, when given a choice between the two dimensions, listeners prefer to make decisions on the basis of one rather than another, even when the two contrasts are equated in terms of perceptual distance in isolation. That is, some acoustic properties may be privileged, at least with respect to their use in distinguishing a given phonetic contrast.

There seem to be at least two or three possible explanations of how such an intrinsic bias might arise. On the one hand, biases might arise as a function of (possibly innate) biological mechanisms, for example, as a consequence of differences in the efficiency of neural systems for processing different kinds of features, e.g., differences in neural systems specialized for processing temporally versus spectrally defined properties, see Zatorre and Belin (2001). Alternatively, such biases might derive from auditory/acoustic interactions between features that result in one feature enhancing the perception of another (Diehl and Kluender, 1989; Kingston and Diehl, 1994) or the two features together contributing to a higher-order, combinatoric perceptual feature (Kingston *et al.*, 2008). Finally, such biases might be explicitly learned, developing through years of experience listening to a language in which linguistically salient differences are more frequently made on the basis of one feature rather than another (a pattern whose origins might itself ultimately have a socio-historical as well as or instead of a psychophysiological basis) (see Holt *et al.* 2001 for discussion). That these kinds of explanations need not be mutually exclusive is supported by recent evidence suggesting that listeners' native language experience affects the efficiency of neural encoding of pitch properties at the brainstem level (Xu *et al.*, 2006).

One of the most recent and thorough discussions of the idea that listeners may be predisposed to use certain acoustic properties rather than others in a categorization task was presented by Holt and Lotto (2006). They trained adult listeners to categorize unfamiliar nonspeech sounds that differed according to two orthogonal dimensions, the center frequency (CF) of the carrier sine wave and the frequency of a modulating sine wave. They found that listeners showed a consistent preference for the CF cue, even when the perceptual distances between the two categories were equal along the two dimensions. This suggests that there may be intrinsic biases favoring the ability to learn (and therefore use) certain acoustic dimensions rather than others (see also Lutfi and Liu, 2007), but it is not known whether this is the case for dimensions that are relevant to perceiving speech sounds.

If English speakers' preference for using VOT over onset F0 in determining a syllable-initial stop-consonant voicing contrast results from a privileged status for VOT, then we would expect VOT to be given more weight than onset F0 when perceiving a voicing contrast even when the perceptual distance between tokens is equalized along the onset F0 and VOT dimensions. Thus, the first goal of the present study is to determine whether VOT and onset F0 exhibit different weighting in a voicing decision when perceptual distance is not a factor. These two commonly studied acoustic correlates of the phonetic voicing contrast were chosen because of the extensive literature on the perception of these two features and because previous research strongly suggests that VOT is more heavily weighted than onset F0 for perceiving the English voicing contrast in syllable-initial stops, yet it is not known whether this pattern still obtains after equating the two distances perceptually.

B. Dimensional integrality

Another consequence of the multidimensionality of speech sounds is that many acoustically independent correlates covary consistently with one another in the speech signal. The covariance of onset F0 and VOT has been argued to arise from a variety of sources. Abramson (1977) and Lisker (1978) suggest that the two features share a common origin in the unfolding of the same laryngeal timing gesture, while Hombert (1978) links the two via aerodynamic demands (higher airflow following the release of voiceless stops leading to a greater onset F0 and longer VOT),[2] In contrast, others ascribe the covariance to perceptual factors. For example, Kingston and Diehl (1994) and Kingston *et al.* (2008) argue that the two cues contribute to the perception of an overarching property of low frequency energy continuing into the stop closure (near short VOT/low onset F0 consonants) or its absence (in long VOT/high onset F0 conso-

nants), while Holt *et al.* (2001) claim that the covariance is learned simply because the two cues are reliably associated in the ambient language (without specifying a basis for this association).

In all cases, however, we might expect covarying cues to be highly integral in the sense of Garner (1974). Listeners who are accustomed to hearing that two cues covary in a consistent manner might be expected to have difficulty ignoring irrelevant variability in one of the cues when making a decision based on the properties of the other, especially if the two cues are integrated into a distinct "intermediate perceptual property" (Kingston *et al.*, 2008). When perceptual distances along the two covarying dimensions are not equal, variability along the more distinctive dimension tends to interfere more with classification along the less distinctive one in a pattern of performance known as asymmetric integrality (see Garner, 1974, 1983; Melara and Mounts, 1994). Thus, in the case of the covarying cues of onset F0 and VOT, if the perceptual distance between long- and short-lag VOT categories is naturally greater than that between falling and rising onset F0 categories, then this would be sufficient to explain the primacy of VOT as a cue to voicing, but artificially equating the perceptual distances along both dimensions should result in a symmetrical pattern of interference.

On the other hand, if VOT is intrinsically more attention demanding than onset F0, then variability in VOT should interfere more with classification according to onset F0 than vice versa. Moreover, this dominance should be maintained even when the perceptual distances between stimuli are equated (that is, even when stimuli are selected such that their perceptual distance is equivalent along each of two dimensions tested in isolation), because trial-to-trial changes along a more attention-demanding dimension should attract attention more than those along a less demanding one (see Tong *et al.*, 2008, for a review of some such cases).

In support of the possibility that VOT may simply be a more attention-demanding dimension of contrast, Gordon *et al.* (1993) argue that VOT is a "stronger" phonetic feature than onset F0, in the sense that VOT is more closely linked to the phenomenal quality of voicing than is onset F0. They suggest that under ideal listening conditions onset F0 is more likely to be ignored as a cue to voicing if VOT is unambiguous than vice versa (cf. Abramson and Lisker, 1985). Moreover, Gordon *et al.* (1993) showed that the primacy of VOT over onset F0 as a cue to stop-consonant voicing was mitigated by attentional demands. Under conditions of high cognitive load, listeners showed a decreased reliance on VOT and a corresponding increase in the relative weight given to onset F0, suggesting that, all else being equal, the use of VOT as a cue to voicing attracts or demands greater attentional commitment than using onset F0. However, in the study of Gordon *et al.* (1993) no attempt was made to equate the perceptual distance along the two dimensions. Thus, the second goal of this study was to investigate the symmetry of dimensional interference between onset F0 and VOT when making a voicing decision after equating perceptual distances along both dimensions. In this case, any observation of asymmetric integrality, such that variability in VOT interferes more with classification according to onset F0 than vice versa, would support the hypothesis that VOT is an intrinsically more attention-demanding dimension of phonetic contrast.

C. Perceptual learning

If, in fact, VOT is a privileged dimension for voicing (as compared to onset F0), then listeners might be expected to be better at learning new categories distinguished in terms of VOT than ones distinguished according to onset F0. A variety of studies (e.g., Holt *et al.*, 2004; Pisoni *et al.* 1982) have shown that listeners are able to learn new VOT-based categories with relatively little training, while Francis and Nusbaum (2002) showed that a few hours of laboratory training with Korean speech stimuli were sufficient to induce English listeners to make use of onset F0. However, due to methodological differences it is difficult to compare results across studies. Thus, the third goal of the present study was to determine whether training to identify categories differing only along one of these two dimensions (VOT or onset F0) would have comparable effects, or whether there would be differences in the effects of training based on the dimension being learned.

D. Enhancement and inhibition

A final question concerned the mechanism or mechanisms by which training affected perception of the two dimensions. A few theories of general perceptual learning (Gibson, 1969; Goldstone, 1994; Nosofsky, 1986) have been applied to perceptual learning of speech, primarily to explain the results of first- and second-language learning (Francis and Nusbaum, 2002; Iverson *et al.*, 2003). According to such theories, category learning requires increasing the similarity of tokens within the same category (acquired similarity), while increasing the perceived differences between tokens in different categories (acquired distinctiveness) (see Liberman, 1957, for what is probably the first application of these terms in speech research, and Jusczyk, 1993, for a comprehensive model of first language acquisition that explicitly incorporates these concepts). Such changes are argued to result from changing the relative weighting of different dimensions: Dimensions that are good at distinguishing categories are given more weight (enhanced), while those that do not differentiate categories well are given less weight (inhibited). Existing research provides tentative support for the hypothesis that both enhancement and inhibition of specific dimensions of contrast may operate in perceptual learning of speech. For example, Francis *et al.* (2000) trained two groups of listeners to use one of two competing cues to syllable-initial stop-consonant place of articulation: The slope of the formant transitions or the spectrum of the burst release. While listeners in the formant-trained condition learned to give increased weight to the formant cue, results from those in the burst-trained group were more suggestive of their having learned to give less weight to the formant cue rather than more weight to the burst cue. However, because the perceptual distance between tokens was not equated across the two cues, we cannot tell whether training caused listeners to adjust the weight given to formant transitions because the

stimuli differed more along this dimension of contrast (formant transitions) or because formant transitions are a privileged cue compared to the spectrum of the burst release. Thus, the final goal of the present study was to provide additional data relevant to determining whether training-related changes in the relative weight given to a specific dimension result from inhibition of the uninformative dimension or enhancement of the more informative one.

E. Summary

In the present investigation listeners were trained to hear a familiar consonantal contrast (voiceless aspirated versus voiceless unaspirated stops, e.g., [p] and [b]) according to either onset F0 or VOT while ignoring variability in the other cue. We used acoustic differences that were within a single category (voiceless aspirated) with the goal of ensuring that our stimuli were located within a region of perceptual space that did not contain any already-known discontinuities in auditory sensitivity such as the well-known discontinuity around 20–30 ms along the VOT dimension (cf. Holt *et al.* 2004) or the probable discontinuity between falling and rising frequency transitions (Schouten, 1985).

We used a variety of training stimuli, incorporating aspects of "high variability" training which has been argued by some researchers to be more effective than other common types of laboratory training (see discussion by Iverson *et al.*, 2005), in an attempt to improve learning over what is often observed in short-term laboratory training studies. We included stimuli produced at a variety of places of articulation of the initial consonant, with a variety of vowels, and produced by two different talkers. However, because the pretest and post-test results we report here derive from stimuli that were identical to (some of) those used in training, we cannot make any strong assumptions about what listeners were actually learning because there is no possibility to measure generalization, e.g., to a novel talker, place of articulation, or vowel context.

We measured the perceptual distance between tokens differing according to these two dimensions both before and after training and compared it to the distribution of selective attention between the two dimensions at the same times. All measurements were made from listeners who exhibited a high degree of success in learning. Our focus is on the performance of these successful learners because we were interested in the effect of *successful learning* on the distribution of weight to acoustic cues. By focusing on learners who showed clear improvement in performance, we also increase the validity of any comparison between the effects of learning observed here and those observed in more natural learning tasks (Francis and Nusbaum, 2002) and in actual cases of native language acquisition (e.g., Iverson *et al.*, 2003). We expected that training would increase perceptual distances along the trained dimension while possibly also decreasing distance along the (task-irrelevant) untrained dimension. Corresponding to these changes, following the results of Melara and Mounts (1994), we expected to see an increase in Garner interference when classifying according to the untrained dimension, and a similar decrease in interference when classifying according to the trained dimension.

II. METHOD

A. Subjects

A total of 42 young adults between the ages of 18 and 36 were initially enrolled in this experiment. All of them were undergraduate or graduate students or staff of Purdue University, or residents of the surrounding community. All participants underwent a standard hearing screening [pure tone audiometry at octave intervals between 500 and 4000 Hz at 20 (500 Hz) or 25 dB HL] and a linguistic background questionnaire designed to identify individuals with strongly monolingual perceptual experience. No applicant was enrolled if they failed the hearing screening, had lived for more than two weeks in a non-English speaking environment, grew up speaking any language other than English, or had lived in a household where the predominant language was anything other than English.

Participants were initially randomly assigned to one of two training conditions, VOT training or onset F0 training. However, as the experiment progressed and it became apparent that the VOT training condition was easier than the F0 condition, more participants were assigned to the onset F0 training group to increase the probability of ending up with relatively balanced numbers of successful learners in both conditions. Of the 42 initial participants, 34 completed all phases of the experiment (producing analyzable data), and 24 of these showed evidence of some learning (improvement of at least five percentage points). In all, 16 of these learners (11 women, 5 men) showed evidence of progressing toward expert perception of the contrast on which they were trained, defined as improvement of at least five percentage points above pretest level as well as a final proportion correct of at least 0.70. There were nine such expert learners in the VOT-trained condition (six women, three men) and seven in the F0-trained condition (five women, two men) (see Sec. III B, below).

B. Design

The goal of this study was to investigate the relationship between changes in perceptual distance and the distribution of selective attention before and after successful training to make phonetic decisions based on one acoustic cue as opposed to another. Thus, in addition to the usual pretest-training-post-test structure commonly used in phonetic training studies (e.g., Francis *et al.*, 2000; Francis and Nusbaum, 2002; Guenther *et al.*, 1999; Guion and Pederson, 2007), three kinds of measures were needed, one to assess degree of learning (in order to identify successful learners), one to determine the distribution of selective attention, and one to evaluate perceptual distance. It was also important that this last measure be obtainable even on the pretest, when listeners were expected to be close to chance when using cues on which they had not been trained. To assess learning, the measure of proportion correct responses was used, calculated over the first and last sessions of training. For measuring the distribution of selective attention, a set of related tasks often referred to as a *Garner paradigm* (Garner, 1974) was used.

Finally, to measure perceptual distance, two quantities were obtained: Sensitivity (d') computed from a speeded target monitoring (STM) task and response time (RT) computed from the baseline component of the Garner paradigm (see below). Sensitivity in a STM task was used in addition to the Garner base line task (which was collected in the course of evaluating selective attention, see below) for two reasons. First, the validity of response-time measures may be less reliable when participants are close to chance, as there will be fewer correct responses on which to base average scores, but the stimuli used in this experiment necessarily sounded quite similar to listeners (prior to training) to increase the likelihood of observing training-related improvement, meaning that performance on the initial Garner task would likely be close to chance. Second, since the primary goal of this study was to compare changes in perceptual distance with changes in selective attention, it was thought desirable to obtain a measure of perceptual distance through methods independent of, though similar in task structure to, the methods used to measure selective attention.

A final aspect of the experimental design that may play a role in interpreting the results is the choice of response categories in the Garner paradigm. In a typical Garner paradigm, stimuli differ along dimensions that are consciously identifiable to listeners, e.g., pitch and loudness, or hue and brightness. In such cases, participants can be instructed to identify stimuli according to a value along either dimension (e.g., is the sound "loud or soft" or "high or low pitched"?). However, in the present case the dimensions are expressly not accessible to conscious processing (Allen *et al.*, 2000). In such cases, researchers frequently first train listeners on novel, arbitrarily labeled categories (e.g., "type 1" versus "type 2"), but this was not an option in the present experiment because one of our research questions involved the effects of training and therefore we did not want to train listeners on the stimuli before we could establish a baseline measure of their performance. Instead, listeners were asked to identify stimuli as belonging to one of two categories (e.g., "B" or "P") when the decision was made along the dimension they were (to be) trained on, or according to one of two alternative categories when the decision was made along the untrained (to be ignored) dimension. The identity of the alternative categories, stressed and unstressed, was chosen based on the correspondence between both VOT and onset F0 with stress in English: Stressed syllables typically exhibit both a higher overall F0 and longer VOT than unstressed syllables, and a sharply falling F0 contour is associated with emphatic stress (as in the final syllable of the response "You don't believe that story, do you?" "Yes, I *do*"). However, listeners were not necessarily expected to be as facile with this classification as with the voicing classification so it was used only for the untrained dimension.

C. Stimuli

Six sets of 100 stimuli varying in two dimensions (onset F0 and VOT) were generated from naturally recorded tokens using PSOLA resynthesis (PRAAT 4.2, Boersma and Weenink, 2006).

1. Recording

Initially, multiple productions of each of the nine syllables [p^hi], [p^ha], [p^hu], [t^hi], [t^ha], [t^hu], [k^hi], [k^ha], and [k^hu] were recorded by one adult male and one adult female native speaker of a Midwestern dialect of American English. Recordings were made to digital audio tape using a hypercardioid microphone (Audio-Technica D1000HE) and digital audio tape-recorder (Sony TCD-D8) in a sound-isolated booth (IAC, model No. 403A), and redigitized to disk for analysis and resynthesis at 22.05 kHz sampling rate and 16 bit quantization using PRAAT 4.2 via a SoundBlaster Live! Sound card on a Dell Optiplex running Windows XP. Speakers recorded multiple instances of three repetitions of each syllable. For example, two or three utterances of [p^ha p^ha p^ha] were recorded by each speaker. Only the second token of each group was digitized to maintain similar intonational properties across tokens. The resulting set of 54 tokens (three repetitions of each of nine syllables by two speakers) was carefully analyzed to identify the acoustically cleanest recording of each syllable. Tokens with a comparatively high degree of line noise or breathiness, irregularities in voicing during vowel production, or other acoustic artifacts that could be compounded by the resynthesis process were discarded. In the end, six tokens were selected for each speaker, creating two mostly overlapping sets (with the lack of complete overlap due to acoustic artifacts in specific recordings). For the female speaker, [p^hi], [p^hu], [t^hi], [t^hu], [k^hi], and [k^ha] were selected, and for the male talker [p^hi], [p^ha], [t^hi], [t^h], [k^hi], and [k^ha]. Stimuli derived from the male [p^ha] tokens were used for testing, and stimuli derived from all tokens (including the male [p^ha]) were used in training.

2. Resynthesis

Starting with each of the 12 base syllables, a set of 100 tokens were resynthesized using the PSOLA methods implemented in PRAAT 4.2, creating a grid varying in ten steps along each of two phonetically relevant acoustic dimensions, onset F0 and VOT, for a total of 1200 tokens (100 tokens for each of 12 starting syllables). Along the VOT dimension, stimuli ranged from 35 to 65 ms VOT in approximately 3 ms steps.[3] Variation in onset F0 ranged from a starting frequency of 1.21 times the starting frequency of the unmodified (base) syllable to 0.91 times (125 Hz for the male [pa]), in steps of about 4 Hz (i.e., for the male [pa] stimulus, the starting frequency ranged from 165 to 125 Hz). All onset F0 contours were linear interpolations starting at the defined initial value and decreasing to the original F0 contour over the first 100 ms of the token (ending at 118 Hz). Thus, all onset F0 contours ranged from sharply falling to nearly flat. There were no rising contours in any stimuli. Slopes ranged from −0.07 Hz/ms (in the shortest VOT, lowest slope stimulus) to −0.47 Hz/ms for the most sharply falling contour.

3. Nomenclature

The goal was to identify four stimuli that differed orthogonally according to two dimensions to perceptually equivalent degrees [forming a square in perceptual space, as

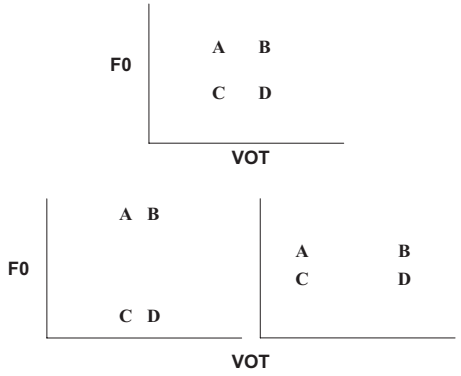

FIG. 1. Hypothetical illustration of changes in perceptual space from equally balanced performance on pretest (1a) to increased attention to VOT/decreased attention to F0 (1b) or decreased attention to VOT/increased attention to F0 (1c). Axes are measured in arbitrary units of perceptual distance.

shown in Fig. 1(a)]. For the purposes of testing and training, stimuli were identified differently to each group, based on the dimension on which each group was trained. For participants in the VOT-trained group, tokens A and C were both treated as exemplars of B while B and D were categorized as P. Conversely, A and B were both considered stressed while C and D were unstressed. In contrast, for participants in the F0-trained group, A and C were both considered unstressed and B and D were stressed, while A and B were labeled as P and C and D were labeled as B.

D. Procedure

Participants completed a total of 11 to 12 sessions, each about an hour in duration, over the course of three to four weeks (one session per day, usually with no more than three days between any two sessions).

The first three sessions and last three sessions constituted the pretest and post-test, respectively, with six sessions of training between them. In the first pretest session participants completed the hearing test, language background questionnaire, and initial assessment of perceptual distance to identify subject-specific, perceptually equal distances along the two dimensions. In the second and third pretest sessions, participants completed the tasks associated with the Garner selective attention paradigm using both male and female stimuli (one talker in each session). The post-test was accomplished in the reverse order of the pretest, but consisted of the same tests (Garner paradigm followed by perceptual distance measurement). When time permitted, the last two sessions of the post-test were conducted on the same day. Training was carried out in the intervening sessions.

1. Perceptual distance measurement (STM)

The goal of this stage of the pretest was to identify four tokens whose pairwise perceptual distances were approximately equal in each of the two dimensions, roughly forming a square in the VOT-by-onset-F0 space, as shown in Fig. 1(a). Sensitivity, d', was used as a measure of perceptual distance because, with listeners expected to be close to chance on the pretest, such a measure would be more informative than response time for correct responses, which might be highly variable due to a high incidence of guessing. Testing always proceeded in the same order. Starting with the B token (step 7 along both the VOT and onset-F0 dimensions, indicating a token close to but not quite prototypical for [p^h]), a corresponding A token was selected having the same onset-F0 value (step 7), but a more [b]-like (shorter) VOT (generally step 3 or 4). Participants then completed a series of eight repetitions of a speeded target monitoring task (STM, see below) using these two stimuli, and sensitivity (d') was calculated as the difference between the z-score transformed proportion of hits and false alarms [$z(H) - z(\mathrm{FA})$ (Macmillan and Creelman, 2004), where hits were counted as correct responses to presented targets, while false alarms were incorrect responses to distractors (nontargets). If the listener's sensitivity to the initial A-B pair was less than 1, a more distant candidate for the A token was selected (e.g., step 2 or 1) and the STM task was repeated. Conversely, if the listener's sensitivity to the initial A-B pair was greater than 1, a closer candidate for the A token (e.g., step 4 or 5) was selected and the STM task was repeated. This process was repeated until either (1) a VOT step value was identified that was approximately 1 d' distant from the B token along the VOT dimension or (2) the perceptual distance between the B token and the most distant possible A token (VOT step 0) was determined. At this point the A token was *fixed* and the selection of a D token began. If the most distant A token was selected (i.e., if the maximum distance between the B and A tokens was still less than 1 d'), then the d' value calculated between this A and the B token was used as the critical value (instead of 1) for the next leg of the square. A similar quasi-iterative process was used to select a D token located approximately the same distance away from the B token along the onset-F0 dimension (typically close to 1 d', but sometimes less if the step-0 A token was used). This process took between one and five repetitions for the AB distance (mean=2.2, SD=0.81) and between one and four repetitions for the BD distance (mean=2.2, SD=0.76). After A and D tokens had been identified through these iterative procedures, a C token was automatically selected having the onset-F0 step value of the D token and the VOT step value of the A token. Once all tokens were selected, the perceptual distances between the remaining adjacent pairs (DC and AC) as well as the diagonals (AD and BC) were computed using the same STM task (see Sec. III). In this way, a set of four tokens were selected that were approximately equidistant in perceptual space for each individual listener. Step values identified in this session were then used for all stimuli, both in testing and training. Note that, since the order of presentation of each pair was the same for all listeners, some effect of order of presentation may have occurred.

The task used to determine d' for a given pair of stimuli was STM. For every pair of tokens, listeners completed one set of eight trials with each trial consisting of a total of 20 stimulus presentations. In each trial, participants were shown a type of sound to monitor for (e.g., B or P for tokens differing only along the trained dimension or stressed or unstressed for tokens differing only along the untrained dimension). The stimulus corresponding to this identifier was considered a target for this trial, while the other stimulus was

considered the distractor. For example, if a member of the VOT-training group was being tested on the distance between the C and D tokens, in a trial specified as monitoring for B, the C token (more [b] like) would be the target while the D token (more [p] like) would be the distractor. If the trial involved monitoring for P then the D token would be the target and the C token would be the distractor. The category identifier (e.g., B) remained on the screen for the duration of the trial. Beginning 1 s after the target identifier appeared on the screen, listeners heard a series of 20 tokens, presented with 1250 ms stimulus onset asynchrony. There were an equal number of target and distractor tokens, and these could appear in any order within the trial with the constraint that a target token could not appear first or last in the trial. Participants were instructed to press a response key every time they heard a syllable starting with the sound shown on the screen and not to respond if the syllable began with a sound different from the symbol shown. They were asked to response as quickly as possible, but also to be as accurate as possible. Responses were scored as hits (responses to targets) or false alarms (responses to distractors) and combined over all eight trials (total of 80 target presentations and 80 Distractors) and used to calculate d'.

Before each trial, listeners were familiarized with the two tokens to be used, and their respective labels for the particular contrast being tested (e.g., for a participant in the VOT-trained group, the A versus B stimulus contrast would be presented as exemplars of B (paired with the A token) and P (paired with the B token). Familiarization consisted of presentation of a stimulus label (e.g., B) with instructions to click on the mouse button in order to hear an example (the A token). After one presentation, listeners were instructed to click the mouse again to hear the sound again. Then the task proceeded to the next stimulus/label pair. Thus, each stimulus was presented a total of 16 times with its associated label in a given block (twice per each of eight trials).

2. Garner paradigm

A complete Garner selective attention paradigm consists of three kinds of tasks, each using stimuli drawn from a set of four stimuli, arranged in a square in perceptual space. The tasks are typically referred to as *baseline, correlation*, and *orthogonal* or *filtering* (Garner, 1974; Pomerantz *et al.* 1989). Each task involves classifying two or four stimuli as exemplars of two categories, e.g., B or P. In this experiment participants completed two base line tasks, two correlation tasks, and one filtering task for each dimension of classification. Because our focus is on Garner interference, only results from the baseline and filtering tasks will be discussed in detail, although responses to some of the stimuli in the correlated condition (specifically, the A and D tokens) are informative with respect to the question of the relative weighting of the two cues in a directly conflicting condition analogous to that used by Francis *et al.* (2000). Moreover, although both male and female voices were used, only results for the male stimuli will be discussed because performance was noticeably better for this talker, especially among the F0-trained listeners. Tasks were blocked by talker (in different sessions) and by dimension: All tasks involving classification by the trained dimension were grouped together, as were all involving classification according to the untrained dimension. Furthermore, the order of labels on the screen (e.g., B and P) and their associated response keys was counterbalanced within blocks for each listener, such that the first half of each block of trials used one order (e.g., B on the left, P on the right) while the second half used the other order. Other than this, tasks were randomized.

TABLE I. Structure of Garner paradigm experiment showing stimuli and tasks for both groups in all conditions.

VOT-trained group			
Trained dimension "Is it B or P?"		Untrained dimension "Is it stressed or unstressed?"	
Task	Stimuli	Task	Stimuli
Base line 1	A, B	Base line 1	A, C
Base line 2	C, D	Base line 2	B, D
Filtering	A, B, C, D	Filtering	A, B, C, D
Correlation 1	A, D	Correlation 1	A, D
Correlation 2	B, C	Correlation 2	B, C
F0-trained group			
Trained dimension "Is it B or P?"		Untrained dimension "Is it stressed or unstressed?"	
Task	Stimuli	Task	Stimuli
Base line 1	A, C	Base line 1	A, B
Base line 2	B, D	Base line 2	C, D
Filtering	A, B, C, D	Filtering	A, B, C, D
Correlation 1	A, D	Correlation 1	A, D
Correlation 2	B, C	Correlation 2	B, C

In each of the baseline and correlation tasks, listeners heard repetitions of only two different stimuli, e.g., the A and B tokens or the A and the C tokens, and classified them according to the appropriate categories by pressing a button on a button box corresponding to the category label shown on that side of the screen. For example, A and B would be classified as B and P, respectively, by participants in the VOT-trained group classifying stimuli along the trained dimension, but as unstressed and stressed by participants in the F0-trained group classifying stimuli along the untrained dimension. In the correlation condition stimuli were classified according to both dimensions. For example, the contrast between A and D would be classified as "B and stressed" versus "P and unstressed" by listeners in the VOT-trained condition, and as "P and unstressed" versus "B and stressed" by listeners in the F0-trained condition. In the filtering condition listeners still made a binary decision, e.g., B or P, but all four stimuli were presented in random order (see Table I for a complete description of the distribution of stimuli in each task).

In the base line and correlated conditions there were a total of 64 trials with each pair of sounds (32 trials per stimulus, in random order within blocks). Response choice location and corresponding button was counterbalanced within each block (e.g., half of the trials showed the order "B" "P" and the other half showed "P" "B" from left to right), for a

total of 128 stimulus presentations for both dimensions of contrast (trained and untrained). In the filtering condition there were also 128 total trials (32 per stimulus) and response choice location was similarly counterbalanced. Before the Garner paradigm began, listeners completed a minisession consisting of two trials of each of the two baseline tasks (in random order). Before every block (practice, each baseline condition, each correlated condition, and the filtering task) listeners were also familiarized with the stimuli and their respective labels to be used in the current block, in the same manner as for the STM task. However, unlike the STM task, familiarization was carried out before each block of the Garner task, not before each trial.

Response times for each correct response were averaged according to Dimension of classification (either trained or untrained) and task (base line, filtering) for each subject, and Garner interference was calculated as (filtering RT—baseline RT) for each dimension. Response times were measured from the beginning of the stimulus and no response times less than 350 ms (the maximum duration of the longest male stimulus) were recorded.

3. Training

The six sessions between the pre-test and post-test consisted of training. In each session, listeners heard six blocks of trials, three with the male voice and three with the female one. Each block of trials consisted of stimuli with a different place of articulation (bilabial, alveolar, and velar). Possible responses were always appropriate to the place of articulation (e.g., P or B for the bilabial blocks, "T" or "D" for the alveolar blocks, and "K" or "G" for the velar blocks). In each block, listeners heard eight different stimuli, presented in random order, ten times each. As in the Garner tasks, the trials in the first and second halves of each block used a different response order left to right. The stimuli consisted of the tokens corresponding to those identified in the initial perceptual distance measurement, but with the appropriate consonant place of articulation and vowel quality for the given block. For example, once a given participant demonstrated roughly equal perceptual distances between four /pa/ stimuli, then in the velar blocks of trials that participant would have heard /ka/ and /ki/ syllables with onset F0 and VOT values corresponding to the same steps along their respective continua.

III. RESULTS

A. Training

Overall, training was successful. Looking at performance on the first and last (sixth) days of training, across all training stimuli (male and female, at all places of articulation and in all vowel contexts included in the experiment), listeners in the VOT group improved from 68% to 81% correct, while those in the F0 group improved from 60% to 67% correct. Results of a repeated measures ANOVA with the two factors of group (VOT trained and F0 trained) and training session (days 1 and 6) showed a significant effect of session, $F(1,32)=4.40$, $p=0.001$, and of group, $F(1,32)=9.31$, $p=0.005$, but no interaction, $F(1,32)=3.18$, $p=0.08$. Planned comparisons of means (significance reported for tests at $p<0.05$ or better) on the pretest showed a significant difference between participants assigned to VOT training and those assigned to F0 training, and this difference remained significant on the post-test. However, both groups improved significantly from day 1 of training to day 6. A t-test of difference scores showed no significant difference between the improvement from day 1 to day 6 for the VOT group (13%) and that shown by the F0-trained group (8%). However, this may be a result of the large amount of variance in changes in performance, since 13 out of the 14 participants (93%) in the VOT-trained group showed an improvement from pretest to post-test, as compared to only 15 out of 20 (75%) in the F0-trained group, despite the equalization of perceptual distance along each dimension on a participant-specific basis. This suggests that listeners who were able to learn the F0 contrast were comparatively few, but showed relatively large improvements, while those who learned the VOT contrast were more common, but did not generally show such extreme improvements.

Because we were interested in understanding the effects of learning (successful training), we restricted subsequent analyses to results only from those participants who both achieved at least 70% correct on the final day of training and showed at least 5% improvement in token identification from the first to the last day of training. Repeating the same analysis on only these 16 participants (7 in the F0 group, 9 in the VOT group) showed the expected significant effect of test, $F(1,14)=69.75$, $p<0.001$, but no effect of group, $F(1,14)=3.23$, $p=0.09$, and no interaction, $F(1,14)=0.10$, $p=0.76$ (Fig. 2). Planned comparisons of means showed again that both groups improved significantly (VOT, from 72% to 88% correct; F0 from 64% to 82% correct), but there was no significant difference between the groups on either the pretest or post-test. This suggests that successful learners from both groups showed comparable improvements in performance along the dimension on which they were trained.

B. Perceptual distance (STM)

Responses to targets in the go/no-go STM task were scored as hits while responses to distractors were scored as false alarms. Perceptual distances between each pair of tokens are shown in Table II. Results of a mixed factorial ANOVA of the pretest distances with the between-groups factor of group (VOT-trained, F0-trained) and within-groups factor of pair (AB, CD, AC, BD, and the diagonals AD and BC) showed a significant effect of pair, $F(5,70)=8.05$, $p<0.001$, but no effect of group, $F(1,14)=3.79$, $p=0.07$, and no interaction, $F(5,70)=0.62$, $p=0.68$. *Post hoc* (Tukey HSD, $p=0.05$) tests showed a significant difference *only* between the pairs that make up the sides of the square (AB, CD, AC, BD) and those making up the diagonals (AD and BC), as Euclidean geometry would predict for a square. There were no significant differences between any two sides of the square, and none between the two diagonals, suggesting that the stimuli selected were perceptually "square" (all sides equal, and both diagonals equal). A similar analysis of the post-test data showed comparable results: A significant

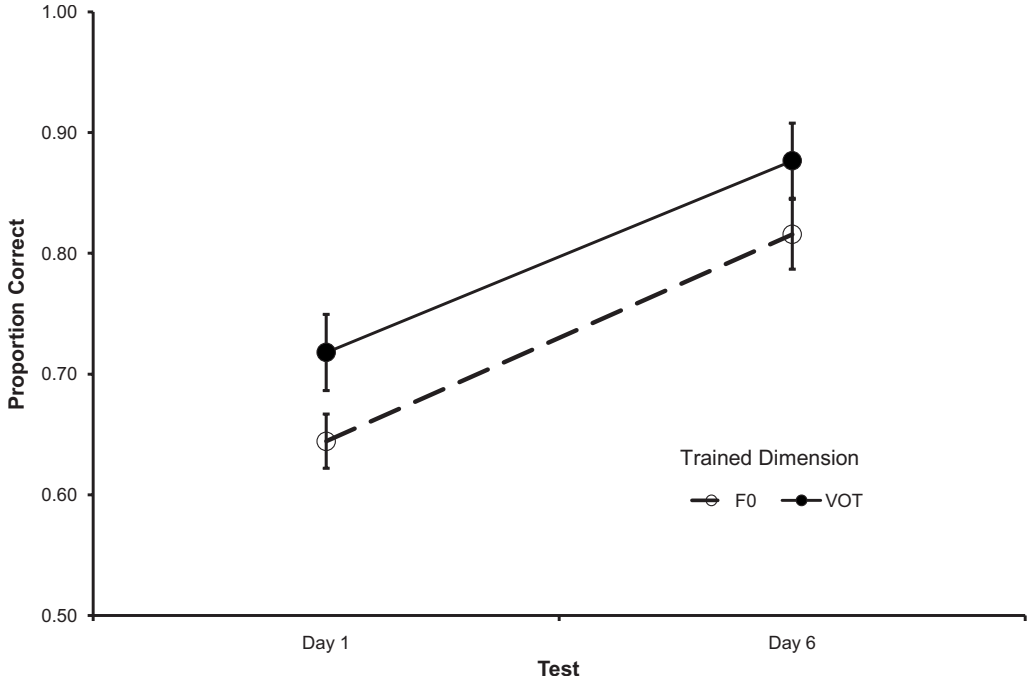

FIG. 2. Proportion of correct consonant identification responses on the first and last days of training for both training groups (successful learners only, see text). Error bars indicate standard error of the mean.

effect of pair, $F(5,70)=9.88$, $p<0.001$, but no effect of group, $F(1,14)=0.18$, $p=0.68$, and no interaction, $F(5,70)=1.78$, $p=0.13$. Again, *post hoc* analyses showed no significant differences between any two sides of the square, and no difference between the two diagonals, although the diagonals were again significantly longer than the sides.

In order to compare performance from pretest to post-test, parallel legs of each square were averaged (e.g., AB and CD were averaged, as were AC and BD) to derive a measure of sensitivity to each dimension for each subject. Results of a mixed factorial ANOVA with between-groups factor of group (VOT-trained, F0-trained) and repeated measures of test (pretest, post-test) and dimension (VOT, onset F0) showed a significant effect of test, $F(1,14)=36.39$, $p<0.001$, but no main effects of group, $F(1,14)=1.02$, $p=0.33$, or of dimension, $F(1,14)=0.30$, $p=0.59$. There was a significant interaction between dimension and group, $F(1,14)=5.35$, $p=0.04$, but no significant interactions between dimension and test, $F(1,14)=0.45$, $p=0.51$, group and test, $F(1,14)=0.29$, $p=0.60$, or between group, dimension and test, $F(1,14)=1.30$, $p=0.27$. Planned comparisons of means (all values reported as significant at $p<0.05$ or better) showed that, for the VOT group, there was a significant increase in sensitivity to the VOT dimension (from a d' of 1.93–3.30) and the F0 dimension (from a d' of 1.51–2.72). Similarly, for the F0-trained group, d' for the VOT dimension increased significantly from 1.28 to 2.51, while for the F0 dimension it increased significantly from 1.29 to 3.12. This suggests that the effect of training on perceptual distance was robust and not constrained to the dimension on which listeners were trained. Overall, these results suggest that the perceptual distances between tokens along each dimension were successfully equated on the pretest, and remained equal after training. Thus, with respect to measures of perceptual distance based on accuracy of speeded target monitoring, training primarily served to increase perceptual distances, and did so to an equivalent degree along both the trained and untrained dimensions.

TABLE II. Perceptual distance, in d' units, fro all pairs of stimuli for both groups on pretest and post-test.

	VOT-trained				F0-trained			
	Pretest		Post-test		Pretest		Post-test	
Pair	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AB	1.59	0.45	2.83	1.65	1.06	0.55	2.20	0.75
CD	2.28	0.84	3.77	1.59	1.49	0.79	2.87	1.81
AC	1.54	0.90	2.88	1.20	1.32	0.70	3.68	1.91
BD	1.48	0.42	2.55	1.15	1.25	0.43	2.56	0.75
AD	2.86	1.59	4.32	1.65	1.74	0.89	4.37	0.82
BC	3.32	1.73	4.52	1.00	2.49	1.26	3.98	1.29

C. Perceptual distance (Garner baseline RT)

Although perceptual sensitivity can be measured in terms of response sensitivity (hit rate and false alarm rate), measures based on response time may be better at differentiating subtle training-related differences between groups. Thus, response times for correct responses in the base line Garner task were averaged for each learner and dimension of classification to provide another measure of perceptual distance between tokens before and after training. Responses made when classifying according to the trained dimension reflect correct responses to the question "is this B or P" while those made when classifying according to the untrained di-

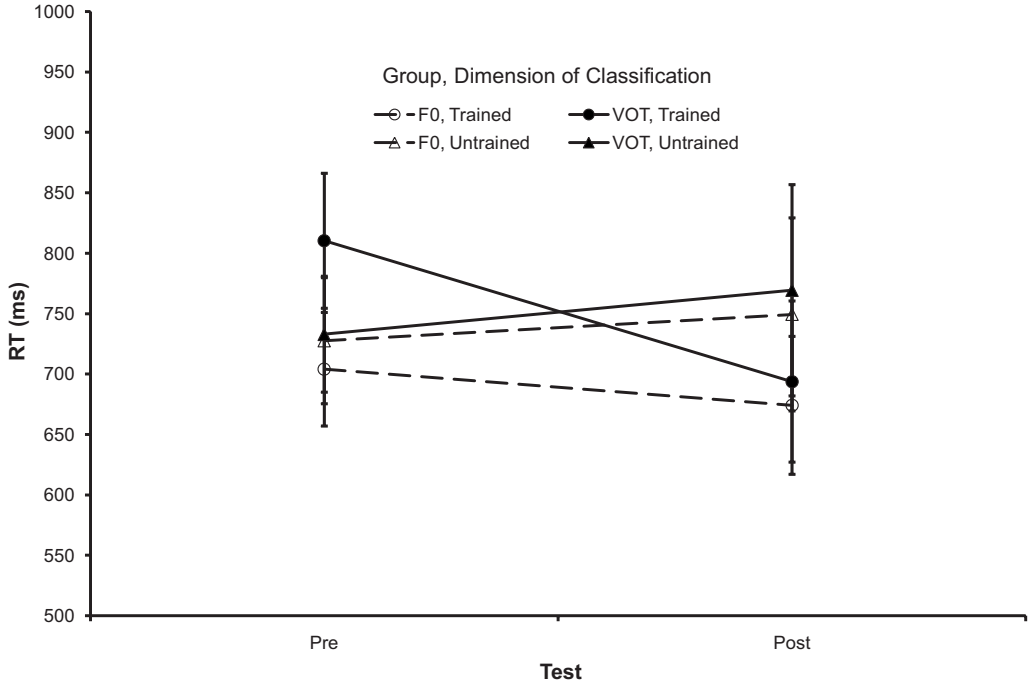

FIG. 3. Pretest and post-test response times on the Garner base line task, classifying stimuli as either [b] or [p] (trained dimension) or "stressed" or "unstressed" (untrained dimension) for both training groups, separated by dimension of classification. Error bars indicate standard error of the mean.

mension reflect response times for classifying according to the other dimension, in response to the question "Is this sound stressed or unstressed?"

A repeated measures ANOVA with one factor between groups (training group, either VOT or onset F0) and two factors within group (test and dimension) showed no significant effects of group, $F(1,14)=0.23$, $p=0.64$, test, $F(1,14)=0.50$, $p=0.49$, or dimension, $F(1,14)=0.53$, $p=0.48$, and no significant interactions between test and group, $F(1,14)=0.33$, $p=0.57$, or between dimension and group, $F(1,14)=0.53$, $p=0.48$. However, the interaction between group, test, and dimension was significant, $F(1,14)=13.35$, $p=0.003$, as shown in Fig. 3. *Post hoc* (Tukey HSD) analysis with a significance threshold of $p=0.05$ showed that the only significant pairwise comparison in the three-way interaction was the 116 ms decrease in baseline response time from pretest (810 ms) to post-test (694 ms) for the VOT-trained group classifying tokens according to the trained (VOT) dimension. The observation that none of the pairwise comparisons for pretest response times showed a significant difference corroborates the findings from the STM task, supporting the claim that stimuli were indeed a perceptual square prior to training. However, the pattern of change in RT, unlike the pattern of change in sensitivity, suggests that only the VOT-trained group showed any appreciable change in perceptual distance between tokens as a result of training, specifically an increase in the distance between tokens along the VOT dimension.

D. Cue weighting

On the pretest, in the correlated task, learners showed no strong evidence in favor of one dimension over another. In the correlated condition involving the A and D tokens stimuli exhibited conflicting values of VOT and onset F0. The A token had a short VOT (similar to a [b]) but a falling F0 contour (like a [p]), while the feature values were reversed for the D token (long VOT like [p] but level F0 onset, more like [b]). Thus, a response of B to the A token or P to the D token would indicate a decision made according to VOT, while a P response to A or a B response to D would indicate a decision made according to onset F0. Overall, learners showed no preference for either cue: 49% of responses to the A token and 51% of those to the D token were consistent with the F0 cue, and this pattern remained even on the post-test (51% and 48%, respectively). This lack of a preference for one cue over another suggests that the bias toward using VOT under normal circumstances (when other cues do not conflict) is not due to something about the VOT dimension *per se*, but rather has to do with the relative size of the interstimulus differences in VOT as compared to those in onset F0.

There was also a very large difference in response patterns between the two training conditions, even on the pretest. The F0-trained group made 88% of pretest and 96% of post-test responses to both the A and D tokens based on onset F0 (responding P and B, respectively), while the VOT group made only 10% and 7% of their responses to the A and D tokens based on F0, respectively (again, responding P to A and B to D). This suggests that the small amount of familiarization that listeners received prior to beginning the pretest was already sufficient to induce them to make phonetic decisions on the basis of the trained rather than the untrained cue. These results suggest, in turn, that listeners' use of a particular cue may be strongly influenced by even short-term experience with a talker or context.

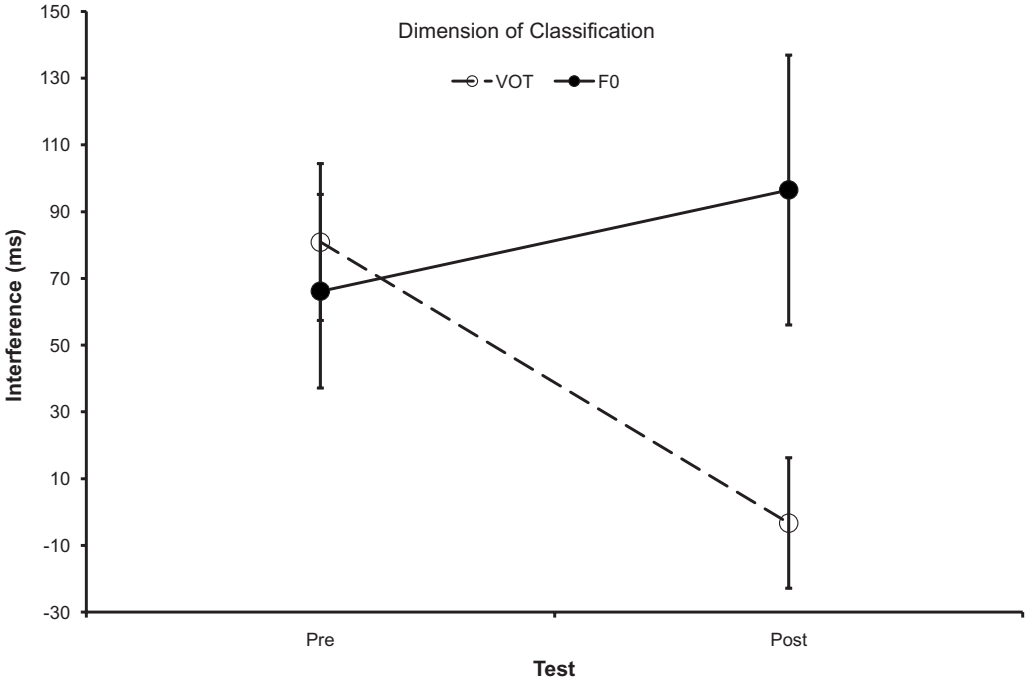

FIG. 4. Garner interference (difference between RT on the Garner filtering task and RT on the Garner base line task, see text for description of tasks) showing significant interaction between test and dimension of classification. Error bars indicate standard error of the mean.

E. Garner interference

Comparison of learners' pretest base line RT with their corresponding filtering RT using a three-way mixed factorial ANOVA with repeated measures of task (baseline, filtering) and dimension of classification (VOT and onset F0), and between-groups factor of training group (VOT and onset F0) showed a significant effect of task, $F(1,14)=9.51$, $p=0.008$, but no effects of group, $F(1,14)=0.54$, $p=0.48$, or dimension, $F(1,14)=2.47$, $p=0.14$, and no interactions. Filtering performance was overall slower (817 ms) than baseline (743 ms) by 74 ms, suggesting that the two dimensions are indeed integral.

Garner interference was computed as the difference in response time between classification according to a given dimension in the filtering task and the average response time for classifying stimuli according to the same dimension in the two baseline tasks using that dimension. These values were submitted to a repeated measures ANOVA with one factor between groups (training group) and two factors within group (test and dimension). Results showed no significant effect of group, $F(1,14)=0.08$, $p=0.78$, test, $F(1,14)=0.62$, $p=0.45$, or dimension, $F(1,14)=1.69$, $p=0.21$, and no interactions between group and test, $F(1,14)=0.08$, $p=0.78$, or group and dimension, $F(1,14)=1.86$, $p=0.19$, and the three-way interaction between test, group, and dimension was not significant, $F(1,14)=1.27$, $p=0.28$. However, there was a significant interaction between test and dimension, $F(1,14)=8.26$, $p=0.01$, suggesting that training had a different effect on the degree of interference of each dimension (Fig. 4). After training, irrelevant variation in F0 no longer interfered with classification according to VOT, but irrelevant variation in VOT continued to interfere with classification according to onset F0.

Although the overall three-way interaction (group by test by dimension) was not significant (Fig. 5), the theoretical basis for the study, namely, the question of whether different kinds of training induce different changes in the processing of the two different dimensions, justified closer examination of some of the contrasts within this interaction. Therefore, a series of planned comparisons were carried out to compare, for each group, the amount of interference for each of the two dimensions on the pretest and on the post-test, as well as the amount of interference for each dimension on the pretest versus the post-test. Significance was set at $p<0.05$. Results showed that, for the F0-trained listeners, there was no significant difference between the degree to which F0 interfered with VOT classification and vice versa on either the pretest or the post-test, and there was no significant difference from pretest to post-test in either the interference of F0 on VOT or vice versa. For the VOT-trained listeners there was no significant difference between VOT or F0 interference on the pretest, but a significant increase from pretest to post-test in interference of VOT on classification according to onset F0 resulted in there being a significant difference on the post-test between the interference of VOT on F0 as compared to vice versa. There were no significant differences in F0 interference from pretest to post-test for this group either.

IV. DISCUSSION

A. Training

Although training can be considered successful for both groups, the degree of learning was unexpectedly low as measured in terms of change in proportion of correct responses from first to last day of training and in terms of the number of trained listeners who exhibited the requisite improvement

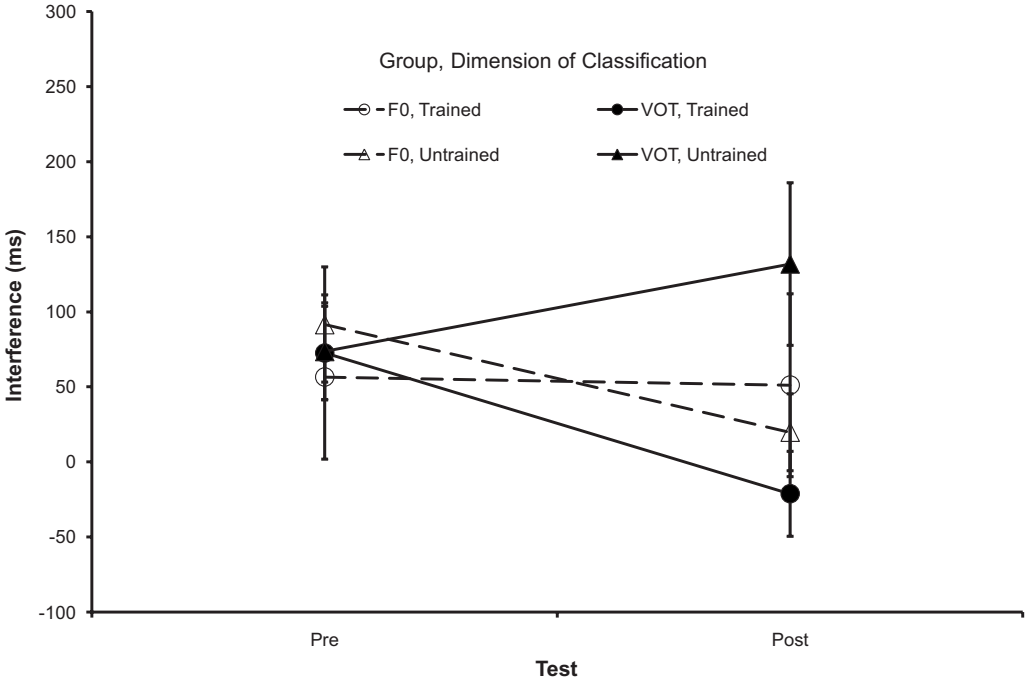

FIG. 5. Differential effects of training on Garner interference (difference between RT on the Garner filtering task and RT on the Garner base line task) for VOT- and F0-trained groups, separated by dimension of classification. Error bars indicate standard error of the mean.

in performance. Previous studies training listeners to develop new categories based on non-native VOT differences (e.g., Holt *et al.*, 2004; Pisoni *et al.* 1982) gave less training and yet showed noticeably better improvement than was found in the present experiment, even for the VOT-trained listeners. Although the training results of Pisoni *et al.* (1982) may have been better than those observed here because of their use of a different location in the VOT continuum (they trained listeners to distinguish between a prevoiced category with negative VOT and a short-lag category), the intended category boundary of experiment 1 of Holt *et al.* (2004), "inconsistent" group, was quite similar to the VOT difference in the present experiment, yet listeners of Holt *et al.* (2004) achieved an identification rate of 90% correct or better within about eight blocks of training (about 380 stimulus presentations).

One possible explanation for the poor rate of learning in the present experiment is that, by using very similar VOT and onset F0 values for all of the training stimuli, regardless of place of articulation (POA), we provided less variability than would be found in natural speech. More significantly, this lack of variability is contrary to the typical correlation between VOT and POA, in which VOT increases as POA moves back in the oral cavity (from bilabial to alveolar to velar) (Lisker and Abramson, 1964). The lack of an expected correspondence of this sort between POA and VOT may have made the additional (non-[pa]) tokens less effective for training, and might conceivably have interfered with learning in some way.

Another major factor that probably contributed significantly to the comparatively low learning rate for listeners in the present experiment is the inconsistent mapping between response category and response button in both testing and training. Although this was done intentionally in an attempt to encourage listeners to develop more abstract categories less closely associated with a specific response key, it almost certainly made the task considerably more difficult. Shiffrin and Schneider (1977) have shown that it is much harder to learn an inconsistent mapping between stimulus and response in which the assignment of stimulus to response changes than a consistent mapping in which stimuli have the same response across trials. Although in the present case the mapping was, at one level, consistent (i.e., shorter VOT values always mapped onto the response B for listeners in the VOT-trained condition), the mapping between the category label B and the response key (left or right) was inconsistent, and this presumably contributed to poorer performance on this task.[4]

B. Perceptual weighting

In the present study, perceptual distance was successfully equated along the two dimensions of VOT and onset F0, as indicated by the results of the pretest STM (d') and Garner base line (RT) tasks. This suggests that the typically observed pattern of using VOT in preference to onset F0 as a cue to voicing in syllable-initial stops (e.g., Abramson and Lisker, 1985; Francis and Nusbaum, 2002; Gordon *et al.*, 1993; Lisker, 1978) can apparently be eliminated at least at the level measurable by discrimination and classification (and at least for tokens that lie within the onset F0 and VOT range of voiceless aspirated stops). In addition, overall performance on the conflicting-cue tokens in the correlated task suggested that listeners showed no *a priori* preference for using VOT over F0, and just a few instances of familiarization were sufficient to induce listeners from both groups to rely heavily on one cue instead of the other. This further supports the hypothesis that preference for VOT is based

strongly on unequal perceptual distance, and does not derive from any special intrinsic property of VOT as a dimension of perceptual contrast.

C. Dimensional integrality

With respect to the question of integrality, results from the Garner interference task on the pretest suggest that the two dimensions of VOT and onset F0 are integral in the sense of Garner (1974). This is consistent with other research on the integrality of speech dimensions (Kingston and Macmillan, 1995; Kingston *et al.*, 1997; Macmillan *et al.*, 1999). Interference was symmetrical on the pretest, such that there was no significant difference in magnitude between the interference of irrelevant variability in onset F0 on classification according to VOT and vice versa, for either of the two groups of learners. This pattern of results is consistent with the hypothesis that any preference for using VOT over onset F0 in classifying voicing in syllable-initial stop consonants derives from unequal perceptual distances along the two dimensions, and not from any preferred quality of VOT. When the perceptual distances were equated along both dimensions in the present experiment, integrality was symmetrical. However, after training, asymmetry increased, at least for the learners in the VOT group, such that there was significantly less interference from irrelevant variability in the untrained dimension (onset F0) on classification according to the trained dimension (VOT) than vice versa. These results (for the VOT-trained listeners), in turn, are consistent with the hypothesis that training served primarily to increase perceptual distance along the trained dimension (VOT). As demonstrated by Melara and Mounts (1994), unequal perceptual distances between tokens along two different dimensions result in increased interference from the larger dimension. Results of the present experiment suggest that, after successfully learning to rely more heavily on VOT and to better ignore onset F0, the perceptual distance between tokens along the VOT dimension was increased with respect to that along the onset F0 dimension for successful VOT-trained listeners, resulting in the observed pattern of increased interference. As discussed below in Sec. IV E other results together suggest that this change resulted primarily from increased distance along the VOT dimension, and not decreased distance along onset F0.

D. Perceptual learning

In this experiment, perceptual distance was calculated in two ways, using d' (sensitivity) in a STM task, and using response time on a Garner speeded classification task. Results were somewhat contradictory, in that the STM task indicated that both groups of learners showed significantly increased perceptual distance along both their untrained and trained dimensions as a result of training, while the classification task indicated that only the VOT-trained group showed an increase in perceptual distance as a result of training, and that occurred only along the trained dimension (see below for a discussion of possible reasons for these differences between monitoring sensitivity and classification response time). At the least, the results from the Garner baseline task lend tentative support to the hypothesis that there may be something special about VOT, as a phonetic cue, that makes it easier to learn than onset F0 (though not easier to use as a cue when perceptual distances are equated): Both groups of listeners were given the same number of trials with the same stimuli, but the VOT-trained group showed, overall, more evidence of stronger learning, including (1) a greater improvement as a result of training (for the entire training group), (2) a greater proportion of listeners showing evidence of learning (greater than five percentage-point increase, with a final score above 70% correct), and (3) the significant changes in Garner interference discussed in the previous section.

While the present results suggest that it may be easier to direct (even) more attention to VOT than to either divert attention from VOT or distribute more attention to onset F0, it is only possible to speculate in a broad manner about possible reasons for such asymmetry in learnability. The most obvious explanation is that American English listeners are simply more used to directing attention to VOT than to onset F0 (cf. Francis and Nusbaum, 2002; Gordon *et al.*, 1993), and thus increasing attention to an already dominant dimension of contrast comes relatively easily. In contrast, inhibiting such a cue may be considerably more difficult, especially since listeners in these studies spend relatively little time in training compared to the amount of time they spend speaking their native language outside the laboratory (where giving greater weight to VOT is clearly a beneficial strategy).

This possibility may be further compounded by the fact that, in testing, listeners were not directed to make judgments about the specific dimensions in question, as would occur in a typical Garner paradigm (e.g., "classify the sounds according to the *pitch* dimension, as either high or low"). Rather, because the dimensions of VOT and onset F0 are not usually thought of as being consciously accessible to untrained listeners, linguistically plausible contrasts were chosen ([b]/[p] for voiced/voiceless, and stressed/unstressed) with the intent that each of these two dimensions should map sufficiently well onto either of the two acoustic cue contrasts (VOT or onset F0). That is, the goal was to use two dimensions such that the mapping between a short VOT stimulus and the response B would be equally acceptable to naïve listeners as that between a short VOT stimulus and the response "unstressed" (and similarly for mappings between shallow onset F0 declines and B and unstressed responses, as well as for long VOT/sharp onset F0 declines and P or "stressed" responses). However, although all expected mappings are plausible *a priori* (stressed syllables do have longer VOT and higher F0 than unstressed ones, and voiced sounds do have shorter VOT and a less negative slope of onset F0 than do voiceless ones), these linguistic dimensions do not, in fact, map equally well onto each respective response for native speakers of English. Not only are English speakers more accustomed to making voicing distinctions based on VOT, not onset F0 (as discussed in the previous paragraph), but they are also more accustomed to making stress distinctions on the basis of F0 than on the basis of VOT. Thus, testing conditions, in terms of the mappings between response items and acoustic dimensions, were much more natural for the

VOT-trained listeners, who were tested with the P/B contrast mapping onto the VOT difference and stressed/unstressed mapping onto onset F0 difference, than for the onset F0-trained listeners, who were tested with the P/B contrast mapping onto the onset F0 difference and stressed/unstressed mapping onto VOT. In other words, our indices of perceptual distance and the distribution of selective attention may be confounded, for the onset F0 group, with experiment design-specific factors, and this might explain why the onset F0 group showed a comparable degree of improvement to the VOT-trained group on the training task (measured in terms of proportion correct identification), but failed to show any evidence of a differential change in the processing of onset F0 as opposed to VOT that might explain this improvement.

On the other hand, it is also possible that there is something intrinsically more learnable about the acoustic properties that comprise VOT as opposed to onset F0 (i.e., an advantage for learning temporal as opposed to spectral contrasts), but to test this hypothesis would require eliminating the bias induced by native language experience, for example, by identifying and training listeners whose native language weighted onset F0 equally with VOT (one such possible example might be Korean, cf. Francis and Nusbaum, 2002). Finally, it may also be noted that training of this sort served primarily to improve the speed with which listeners were able to make a decision, and such an improvement was disproportionately advantageous for decisions based on VOT which is fundamentally temporal in nature and occurs earlier in the syllable, as opposed to onset F0 which involves both spectral and temporal properties and occurs later in the syllable.

E. Enhancement versus inhibition

Although the two methods used to measure perceptual distance (sensitivity in speeded target monitoring versus response time in speeded classification) provided somewhat discrepant results (see below), it is important to note that both methods provided strong evidence that training served only to increase the perceptual distance between tokens (acquired distinctiveness), not to decrease it (acquired similarity). Only the VOT group showed a change in interference, and this was only in terms of the decrease in interference of the untrained on the trained. The (expected) corresponding increase in interference of the trained on the untrained was not significant, although the trend was definitely in the expected direction. Given that the dimensions of VOT and onset F0 are highly integral, these results are entirely consistent with results from previous research. In particular, Goldstone (1994) also found evidence for increased perceptual distance along a variety of trained dimensions in a visual category learning experiment, but only found evidence of decreased perceptual distance along a to-be-ignored dimension when the two dimensions were perceptually separable in the sense of Garner (1974). Indeed, cases of true acquired similarity seem to be relatively rare in the perceptual learning literature [cf. Guenther *et al.* (1999) for discussion, and Francis and Nusbaum, (2002), for an example of acquired similarity with more natural stimuli].

There are at least two ways to characterize the difference between acquired similarity and acquired distinctiveness. Iverson and co-workers (Iverson and Kuhl, 2000; Iverson *et al.*, 2003) have argued that acquired similarity arises from properties of the statistical distribution of input stimuli in perceptual space in a manner independent of attention, while acquired distinctiveness results from the operation of an attentionally demanding process. Although there is now evidence that even passive statistical learning depends on the availability of attentional resources (Toro *et al.* 2005), there is also evidence that the development of acquired similarity can be facilitated by certain distributional properties of the training stimuli. Thus, the Iverson argument may still be valid, despite the almost certain involvement of attention in the process of phonetic cue learning. In support of a role for distributional factors, Guenther *et al.* (1999) found that in order to induce increased similarity, it was necessary to provide not only categorization training (as in the present experiment) but also multiple exemplars of each category. They argued that experience with multiple exemplars encouraged listeners to ignore small (noncategorical) differences between stimuli within a single category, an effect impossible to achieve when training with only a single exemplar [see also Iverson *et al.* (2005) for similar arguments related to a test of the efficacy of high variability training].

On the other hand, Goldstone (1994) and Francis and Nusbaum (2002) argued that the processes of acquired distinctiveness and acquired similarity may be employed at different stages in the learning process, and/or under different conditions of stimulus properties. In cases such as the present experiment and those of Goldstone (1994) in which stimuli are perceptually highly similar (located within a single native category in the present case, or within one (just noticeable difference) JND of one another in the Goldstone case), acquired distinctiveness is the most effective strategy for significantly improving categorization quickly. In contrast, under conditions in which stimuli are already relatively easy to categorize (e.g., certain contrasts in the Korean stimuli used by Francis and Nusbaum, 2002), acquired similarity, especially along an irrelevant dimension of contrast, leads to a more significant improvement in categorization that would simply further increase the already salient difference between the two categories along an already contrastive dimension.

Of course, the two accounts are not necessarily mutually exclusive, in the sense that the presence of multiple exemplars within each category increases the probability that variance within the category is relatively high, which in turn increases the likely benefit of applying a process of acquired similarity to reduce within-category variability. In the present case, however, listeners were trained with multiple exemplars, but these exemplars were acoustically extremely similar to the test stimuli along the critical dimensions of onset F0 and VOT, and yet the two categories represented by these exemplars (and by the test stimuli) were extremely close to one another in perceptual space. Thus, in this case, although listeners received multiple training exemplars, one might argue that they were not distributed in a manner that would be expected to promote acquired similarity on the basis of either

of these two hypotheses. The distribution of training exemplars was not sufficiently broad to engage a Guenther/Iverson type of mechanism, and the overall similarity of the two categories was sufficiently great to engage a mechanism of acquired distinctiveness over one of acquired similarity in a Francis/Goldstone type of model. Further research is clearly necessary to explore the basis for these two kinds of processes.

F. Differences between monitoring sensitivity and classification response time

One curious finding in the present results is the apparent disagreement between the two measures of perceptual distance employed, sensitivity in speeded target monitoring and response time in speeded classification. While the sensitivity results indicated that listeners in both groups showed equivalently increased perceptual distances along both their trained and untrained dimensions of contrast, the response-time data suggested that only the VOT-trained listeners showed a change in perceptual distance, and this increase occurred only along VOT, the dimension on which they were trained.

This finding is particularly curious given the commonly accepted assumption that response time and accuracy tasks are assumed to measure more or less the same thing (perceptual distance between tokens). Ashby and Maddox (1994) discuss the widespread nature of this assumption as they develop an explicit model relating RT performance to perceptual distance between tokens and decision (category) boundaries, based on general recognition theory (GRT) (Ashby and Townsend, 1986). Specifically, they propose that RT should decrease monotonically as a function of the perceptual distance between the stimulus and the decision bound. Furthermore, the GRT as well as other theories of similar phenomena (e.g., Luce, 1986) clearly demonstrate that difficult discriminations are associated with longer response times. Thus, we have every reason to expect a correspondence between RT and accuracy measures: As stimuli become more distant from one another in perceptual space, they should become both easier to identify (in the STM task) and correct identifications should be faster (in the Garner base line task). However, it is possible that, in the present case, specific details of the experiment design unintentionally predisposed listeners to treat the two tasks differently with respect to the type of memory or attentional mechanisms they employed, resulting in a divergence between the results of the two tasks.

One potentially important difference between the two tasks in the present experiment is that, in the STM task, listeners received much more frequent familiarization with exemplars of the two categories they using than they did on the classification task. In the STM task, listeners heard two presentations of each of the two stimuli in a given trial (e.g., the A and B tokens), accompanied by visual presentation of their associated category label, before every trial. On the other hand, in the classification task, listeners were familiarized with the stimulus-symbol pairing only three times, once before each *block* of trials (baseline, correlated, and filtering). Thus, performance in the STM task may better reflect listeners' ability to compare each test stimulus with short-term memory traces of the familiarization stimuli, while performance on the Garner base line task better reflects listeners' ability to compare test stimuli with long(er)-term category representations [see Xu *et al.* (2006) for a model of memory for phonetic categorization].

Macmillan (1987) distinguishes between sensory or trace and context modes of processing. In the trace mode, processing is dominated by comparison of (temporary) sensory traces of stimuli, while in context coding processing involves comparison between sensory traces of stimuli and (longer-term) perceptual anchors, including category representations. In this sense, the different familiarization protocols for the two types of tasks may have encouraged a greater degree of reliance on sensory coding in the STM task and on context coding in the classification (Garner base line) task. That is, performance measured in terms of accuracy on the STM task may serve mainly to indicate listeners' ability to retain and make use of short-term memory traces of the familiarization stimuli. As listeners learned which properties of the signal (VOT and onset F0) varied across the training stimuli, they may have become better able to encode and retrieve these properties as short-term memory traces (i.e., when exposed to the tokens during familiarization). Since both properties varied equally across the training set, listeners showed an equal degree of improvement in encoding and retrieving memory traces of these properties.

On the other hand, RT performance on the Garner base line task may better reflect listeners' ability to access stored long-term representations of phonetic categories (context coding). It has been argued that perceptual learning based on categorization training (as used here) primarily affects categorization at the level of context coding (Guenther *et al.*, 1999). According to this hypothesis, training was successful in changing the long-term representations of the categories that listeners were learning (e.g., B versus P), but this only became obvious in the Garner baseline (RT) task because there was sufficient time between the presentation of the familiarization stimuli and the actual test trials that listeners were not able to rely solely on trace memories of the familiarization stimuli and instead had to depend on their long-term memories of the different (learned) category representations. Thus, it may be argued that the results of the Garner base line task are more indicative of the overall phonetic consequences of this kind of training than are those of the STM task, because they better reflect changes in listeners' attention to features encoded in long-term memory representations of the learned categories, while the results of the STM task reflect instead an increase in overall sensitivity to those acoustic properties that varied during training as a result of increased attention to the speech signal under conditions of higher uncertainty (Nusbaum and Magnuson, 1997; Nusbaum and Schwab, 1986; Wong *et al.*, 2004).

G. The role of attention in phonetic learning

Gordon *et al.* (1993) showed that, under conditions of (comparatively) unlimited attentional load, American English listeners gave more weight to VOT than to onset F0 in a voicing decision. In contrast, under conditions of more limited attentional availability, listeners showed a greater re-

duction in the weight given to VOT than in that given to onset F0. They argued that weak acoustic cues (e.g., onset F0) require comparatively little attention to make their full contribution to a phonetic decision (thus benefiting little from an increased availability of attention), while stronger cues (e.g., VOT) benefit more from increased availability of attentional resources. We elaborate on this hypothesis by proposing that using *any* cue requires some commitment of attention, but that attention is allocated dynamically depending on the current diagnosticity of specific cues. Under normal circumstances those cues that have proven to be most diagnostic (e.g., over the course of prior experience) receive the lion's share. Under conditions of limited attentional availability, the proportion of capacity devoted to each cue is reduced proportionally, with strong cues continuing to receive proportionally more of the smaller pool of available resources. In new contexts or under conditions of uncertainty (i.e., multiple talkers, high noise, etc.), the distribution of attention to individual cues may vary as the speech perception mechanism begins to seek out cues that are potentially more diagnostic under those conditions (Nusbaum and Magnuson, 1997; Nusbaum and Schwab, 1986; Wong *et al.*, 2004). Such reallocation may result in a more even distribution of resources across cues as attention is withdrawn from cues that are typically stronger but fail to be sufficiently diagnostic in the present context, and reallocated toward cues that, though typically weaker, might potentially be more diagnostic in the present case.

In this dynamic redistribution of attention we see a reconciliation between the effects of training and the effects of experimental task observed in the present experiment. On the one hand, perceptual training may alter the base line distribution of attention to specific cues, increasing the weight given to cues that are sufficient for identifying the newly learned categories, and reducing that given to less diagnostic cues. That it does so preferentially for VOT and less so for onset F0 suggests that there is something special about VOT, at least as a cue to the perception of syllable-initial stop-consonant voicing by native speakers of English. On the other hand, frequent presentations of representative stimuli differing along two dimensions (as in the STM task) may encourage listeners to maintain a high level of attention to both cues to facilitate the use of trace coding. Thus, the ability of training to accomplish the redistribution of attention among acoustic cues may only become obvious under conditions in which listeners are not constantly reminded of the multiple dimensions (diagnostic and nondiagnostic) along which stimuli differ, and instead are forced to focus on stimulus differences that have been encoded in the long-term mental representations of the learned categories.

Ultimately, this perspective is compatible with Kuhl's *neural commitment* theory (Kuhl *et al.*, 2006), in the sense that English listeners appear to have committed to VOT to a greater degree than to onset F0 (at least as a cue to the phonetic property of voicing in syllable-initial stops), and reducing that commitment, or increasing their commitment to onset F0, seems to require more training, or different kinds of training, than we have employed here. Whether this commitment derives from innate differences in the neural systems that process VOT as compared to onset F0, or from experience-dependent development of such systems is a question beyond the scope of the present paper. However, by considering such neural commitment in terms of the distribution of attentional resources we are able to link the role of attention in perceptual learning (Guion and Pederson, 2007; Strange, 2006) to processes of online speech perception (Gordon *et al.*, 1993), making a connection that is obviously necessary, but thus far only occasionally discussed (Nusbaum and Goodman, 1994; Stevens *et al.*, 2006; Toro *et al.*, 2005).

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institute on Deafness and other Communication Disorders (NIH-NIDCD R03DC006811) to A.L.F. We would like to thank Bob Melara, Howard Nusbaum, John Kingston, and an anonymous reviewer for suggestions on earlier drafts of this article. Some of these results were presented at the fourth Joint Meeting of the Acoustical Society of American and the Acoustical Society of Japan, Honolulu, HI, November 28–December 2, 2006.

[1]Although the dimension of VOT has been explored in considerable depth, the dimension of onset F0 is less well investigated, and to our knowledge there are no studies that provide quantitative data on listeners' sensitivity to onset F0 differences comparable to the wealth of information available regarding VOT (see Holt *et al.*, 2004 for discussion).

[2]Note that subsequent research (e.g., Löfqvist *et al.*, 1989) supports a physiological origin of the onset F0 property of stop consonants in the degree of tension of the cricothyroid muscle, suggesting that there is no direct physiological link between onset F0 and VOT cues. This physiological dissociation is further supported by the patterning of these two cues in three-way stop consonant systems such as that of Korean and Thai, in which stop categories are distinguished by independent onset F0 and VOT properties (Thai: Gandour, 1974; Korean: Francis and Nusbaum, 2002; see Francis *et al.*, 2006 for discussion).

[3]While step size was maintained as closely as possible across tokens and talkers, when specific values are given here they refer to the test stimuli based on [p^ha]. Other stimuli varied slightly from these specific values to preserve some degree of interstimulus variability, but never by more than 5 ms or two percentage points (for frequency modifications) from the values given here.

[4]In an ongoing study using nonspeech sounds in a similar testing/training paradigm, we have found that simply eliminating this inconsistent mapping between response label and response key improves learning considerably both in terms of the number of listeners who are able to reach criterion, and in terms of the magnitude of the overall change in proportion correct identification from the first to the last day of training.

Abramson, A. S. (**1977**). "Laryngeal timing in consonant distinctions," Phonetica **34**, 295–303.

Abramson, A. S., and Lisker, L. (**1970**). "Discrimination along the voicing continuum: Cross-language tests," Proceedings of the 6th International Congress on Phonetic Science, Prague, **1967**, Academia, Prague, pp. 569–573.

Abramson, A. S., and Lisker, L. (**1985**). "Relative power of cues: F0 shift versus voice timing," in *Linguistic Phonetics*, edited by V. Fromkin (Academic New York), pp. 25–33.

Allen, J., Kraus, N., and Bradlow, A. (**2000**).. "Neural representation of consciously imperceptible speech sound differences," Percept. Psychophys. **62**, 1383–1393.

Ashby, F. G., and Maddox, W. T. (**1994**). "A response time theory of separability and integrality in speeded classification," J. Math. Psychol. **38**, 423–466.

Ashby, F. G., and Townsend, J. T. (**1986**). "Varieties of perceptual independence," Psychol. Rev. **93**, 154–179.

Boersma, P., and Weenink, D. (**2006**). Praat: doing phonetics by computer (Version 4.2) (computer program). http://www.praat.org/ (last accessed March 17, **2008**).
Diehl, R. L., and Kluender, K. R. (**1989**). "On the objects of speech perception," Ecological Psychol. **1**, 121–144.
Francis, A. L., and Nusbaum, H. C. (**2002**). "Selective attention and the acquisition of new phonetic categories," J. Exp. Psychol. Hum. Percept. Perform. **28**, 349–366.
Francis, A. L., Baldwin, K., and Nusbaum, H. C. (**2000**). "Effects of training on attention to acoustic cues," Percept. Psychophys. **62**, 1668–1680.
Francis, A. L., Ciocca, V., Wong, V. K. M., and Chan, J. K. L. (**2006**). "Is fundamental frequency a cue to aspiration in initial stops?," J. Acoust. Soc. Am. **120**, 2884–2895.
Gandour, J. (**1974**). "Consonant types and tone in Siamese," J. Phonetics **2**, 337–350.
Garner, W. R. (**1974**). *The Processing of Information and Structure* (Erlbaum, Hillsdale, NJ).
Garner, W. R. (**1983**). "Asymmetric interactions of stimulus dimensions in perceptual information processing," in *Perception, Cognition, and Development: Interactional Analyses*, edited by T. J. Tighe and B. E. Shepp (Erlbaum, Hillsdale, NJ), pp. 1–37.
Gibson, E. J. (**1969**). *Principles of Perceptual Learning and Development* (Appleton-Century-Crofts, New York).
Goldstone, R. (**1994**). "Influences of categorization on perceptual discrimination," J. Exp. Psychol. Gen. **123**, 178–200.
Gordon, P. C., Eberhardt, J. L., and Rueckl, J. G. (**1993**). "Attentional modulation of the phonetic significance of acoustic cues," Cogn. Psychol. **25**, 1–42.
Guenther, F. H., Husain, F. T., Cohen, M. A., and Shinn-Cunningham, B. G. (**1999**). "Effects of categorization and discrimination training on auditory perceptual space," J. Acoust. Soc. Am. **106**, 2900–2912.
Guion, S. G., and Pederson, E. (**2007**). "Investigating the role of attention in phonetic learning," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, edited by O.-S. Bohn and M. Munro (Benjamins, Amsterdam), pp. 57–77.
Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as a voicing cue," J. Acoust. Soc. Am. **47**, 613–617.
Haggard, M. P., Summerfield, Q., and Roberts, M. (**1981**). "Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F0 cues in the voiced-voiceless distinction," J. Phonetics **9**, 49–62.
Holt, L. L., and Lotto, A. J. (**2006**). "Cue weighting in auditory categorization: Implications for first and second language acquisition," J. Acoust. Soc. Am. **119**, 3059–3071.
Holt, L. L., Lotto, A. J., and Diehl, R. L. (**2004**). "Auditory discontinuities interact with categorization: Implications for speech perception," J. Acoust. Soc. Am. **116**, 1763–1773.
Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2001**). "Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?," J. Acoust. Soc. Am. **109**, 764–774.
Hombert, J. M. (**1978**). "Consonant types, vowel quality, and tone," in *Tone: A Linguistic Survey*, edited by V. A. Fromkin (Academic, New York), pp. 77–111.
Iverson, P., and Kuhl, P. K. (**2000**). "Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism?," Percept. Psychophys. **62**, 874–886.
Iverson, P., Hazan, V., and Bannister, K. (**2005**). "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults," J. Acoust. Soc. Am. **118**, 3267–3278.
Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (**2003**). "A perceptual interference account of acquisition difficulties for non-native phonemes," Cognition **87**, B47–B57.
Jusczyk, P. W. (**1993**). "From general to language-specific capacities: The WRAPSA model of how speech perception develops," J. Phonetics **21**, 3–28.
Kingston, J., and Diehl, R. L. (**1994**). "Phonetic knowledge," Language **70**, 419–494.
Kingston, J., Diehl, R. L., Kirk, C. J., and Castleman, W. A. (**2008**). "On the internal perceptual structure of distinctive features: The [voice] contrast," J. Phonetics **36**, 28–54.
Kingston, J., and Macmillan, N. A. (**1995**). "Integrality of nasalization and F_1 in vowels in isolation and before oral and nasal consonants: A detection-theoretic application of the Garner paradigm," J. Acoust. Soc. Am. **97**, 1261–1285.
Kingston, J., Macmillan, N. A., Dickey, L. W., Thorburn, R., and Bartels, C. (**1997**). "Integrality in the perception of tongue root position and voice quality in vowels," J. Acoust. Soc. Am. **101**, 1696–1709.
Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (**2006**). "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months," Dev. Sci. **9**, F13–F21.
Liberman, A. M. (**1957**). "Some results of research on speech perception," J. Acoust. Soc. Am. **29**, 117–123.
Lisker, L. (**1978**). "In qualified defense of VOT," Lang Speech **21**, 375–383.
Lisker, L. (**1986**). ""Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Lang Speech **29**, 3–11.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.
Löfqvist, A., Baer, T., McGarr, N. S., and Seider Story, R. (**1989**). "The cricothyroid muscle in voicing control," J. Acoust. Soc. Am. **85**, 1314–1321.
Luce, R. D. (**1986**). *Response Times: Their Role in Inferring Elementary Mental Organization* (Oxford University Press, Oxford).
Lutfi, R. A., and Liu, C.-J. (**2007**). "Individual differences in source identification from synthesized impact sounds," J. Acoust. Soc. Am. **122**, 1017–1028.
Macmillan, N. A. (**1987**). "Beyond the categorical/continuous distinction: A psychophysical approach to processing modes," in *Categorical Perception*, edited by S. Harnad (Cambridge University Press, New York), pp. 53–85.
Macmillan, N. A., and Creelman, C. D. (**2004**). *Detection Theory: A User's Guide*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ).
Macmillan, N. A., Kingston, J., Thorburn, R., Dickey, L. W., and Bartels, C. (**1999**). "Integrality of nasalization and F1. II. Basic sensitivity and phonetic labeling measure distinct sensory and decision-rule interactions," J. Acoust. Soc. Am. **106**, 2913–2932.
Massaro, D. W., and Cohen, M. M. (**1976**). "The contribution of fundamental frequency and voice onset times to the /zi/-/si/ distinction," J. Acoust. Soc. Am. **60**, 704–717.
Massaro, D. W., and Cohen, M. M. (**1977**). "Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction," Percept. Psychophys. **22**, 373–382.
Melara, R. D., and Mounts, J. R. W. (**1994**). "Contextual influences on interactive processing: Effects of discriminability, quantity, and uncertainty," Percept. Psychophys. **56**, 73–90.
Nosofsky, R. M. (**1986**). "Attention, similarity, and the identification-categorization relationship." J. Exp. Psychol. Gen. **115**, 39–57.
Nusbaum, H. C., and Goodman, J. C. (**1994**). "Learning to hear speech as spoken language," in *The Development of Speech Perception*, edited by J. C. Goodman and H. C. Nusbaum, (MIT Press, Cambridge, MA), pp. 299–338.
Nusbaum, H. C., and Magnuson, J. (**1997**). "Talker normalization: Phonetic constancy as a cognitive process," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, (Academic, San Diego, CA), pp. 109–132.
Nusbaum, H. C., and Schwab, E. C. (**1986**). "The role of attention and active processing in speech perception," in *Pattern Recognition by Humans and machines*, edited by E. C. Schwab and H. C. Nusbaum (Academic, San Diego), Vol. **1**, pp. 113–157.
Pisoni, D. B., Aslin, R. N., Perey, A. J., and Hennessy, B. L. (**1982**). "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants," J. Exp. Psychol. Hum. Percept. Perform. **8**, 297–314.
Pomerantz, J. R., Pristach, E. A., and Carson, C. E. (**1989**). "Attention and object perception," in *Object Perception: Structure and Process*, edited by B. Shepp and S. Ballesteros (Lawrence Erlbaum Associates, Hillsdale, NJ), pp. 53–89.
Raphael, L. J. (**2005**). "Acoustic cues to the perception of segmental phonemes," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 182–206.
Repp, B. H. (**1979**). "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," Lang Speech **22**, 173–189.
Schouten, M. E. (**1985**). "Identification and discrimination of sweep tones," Percept. Psychophys. **37**, 369–376.
Shiffrin, R. M., and Schneider, W. (**1977**). "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory," Psychol. Rev. **84**, 127–190.
Stevens, K. N., and Klatt, D. H. (**1974**). "Role of formant transitions in the voiced-voiceless distinction for stops," J. Acoust. Soc. Am. **55**(3), 653–659.

Stevens, C., Sanders, L., and Neville, H. (**2006**). "Neurophysiological evidence for selective auditory attention deficits in children with specific language impairment," Brain Res. **1111**, 143–152.
Strange, W. (**2006**). "Second-language speech perception: The modification of automatic perceptual routines. Paper presented at the Fourth Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, November-December, 2006, Honolulu, HI. [Abstract]," J. Acoust. Soc. Am. **120**, 3137.
Tong, Y., Francis, A. L., and Gandour, J. T. (**2008**), "Processing dependencies between segmental and suprasegmental features in Mandarin Chinese," Lang. Cognit. Processes **23**, 689–708.
Toro, J. M., Sinnett, S., and Soto-Faraco, S. (**2005**). "Speech segmentation by statistical learning depends on attention," Cognition **97**, B25–B34.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "F0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**, 2152–2159.
Wong, P. C. M., Nusbaum, H. C., and Small, S. L. (**2004**). "Neural bases of talker normalization," J. Cogn Neurosci. **16**, 1173–1184.
Xu, Y., Gandour, J. T., and Francis, A. L. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am. **120**, 1063–1074.
Xu, Y., Krishnan, A., and Gandour, J. (**2006**). "Specificity of experience-dependent pitch representation in the brainstem," NeuroReport **17**, 1601–1605.
Zatorre, R., and Belin, P. (**2001**). "Spectral and temporal processing in human auditory cortex," Cereb. Cortex **11**, 946–953.