Training English listeners to perceive phonemic length contrasts in Japanese[a)]

Keiichi Tajima[b)]
Department of Psychology, Hosei University, 2-17-1 Fujimi, Chiyoda-ku, Tokyo 102-8160, Japan

Hiroaki Kato
ATR Cognitive Information Science Laboratories/National Institute of Information and Communications Technology, 2-2-2 Hikaridai, Seika-cho, Kyoto 619-0288, Japan

Amanda Rothwell
School of Kinesiology, The University of Western Ontario, London, Ontario N6A 3K7, Canada

Reiko Akahane-Yamada
ATR Cognitive Information Science Laboratories, 2-2-2 Hikaridai, Seika-cho, Kyoto 619-0288, Japan and Graduate School of Cultural Studies and Human Science, Kobe University, 1-2-1 Tsurukabuto, Nada-ku, Kobe 657-8501, Japan

Kevin G. Munhall
Department of Psychology and Department of Otolaryngology, Queen's University, Humphrey Hall, 62 Arch Street, Kingston, Ontario K7L 3N6, Canada

(Received 17 July 2007; revised 10 October 2007; accepted 11 October 2007)

The present study investigated the extent to which native English listeners' perception of Japanese length contrasts can be modified with perceptual training, and how their performance is affected by factors that influence segment duration, which is a primary correlate of Japanese length contrasts. Listeners were trained in a minimal-pair identification paradigm with feedback, using isolated words contrasting in vowel length, produced at a normal speaking rate. Experiment 1 tested listeners using stimuli varying in speaking rate, presentation context (in isolation versus embedded in carrier sentences), and type of length contrast. Experiment 2 examined whether performance varied by the position of the contrast within the word, and by whether the test talkers were professionally trained or not. Results did not show that trained listeners improved overall performance to a greater extent than untrained control participants. Training improved perception of trained contrast types, generalized to nonprofessional talkers' productions, and improved performance in difficult within-word positions. However, training did not enable listeners to cope with speaking rate variation, and did not generalize to untrained contrast types. These results suggest that perceptual training improves non-native listeners' perception of Japanese length contrasts only to a limited extent. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804942]

PACS number(s): 43.71.Hw, 43.71.Es [PEI] Pages: 397–413

I. INTRODUCTION

In Japanese, vowel and consonant length can be used phonemically to distinguish words. The primary acoustic correlate and perceptual cue to such length contrasts are said to be the duration of the vowel or consonant (Fujisaki *et al.*, 1975; Uchida, 1998). Segment duration, however, is affected by numerous factors, including inherent duration, neighboring segments, position within a word or phrase, length of the word or phrase, emphasis or semantic novelty, and speaking rate (e.g., Klatt, 1976; Sagisaka and Tohkura, 1984; Takeda *et al.*, 1989). In fact, the duration of phonemically short and long segments produced across various speaking rates overlap considerably (Hirata, 2004a; Hirata and Whiton, 2005), implying that perceptual judgment of phonemic length cannot be made simply based on absolute segment duration, but must be made in relation to the context in which the segment appears. Such complex behavior of segment duration potentially makes phonemic length extremely difficult for second-language (L2) learners to learn, particularly for native speakers of a language that does not use segment duration phonemically, such as English. Native English speakers have in fact been shown to have difficulty learning to perceive Japanese length contrasts (Yamada *et al.*, 1994; Oguma, 2000; Toda, 2003), but the nature of their difficulty in relation to the various contextual factors has not been thoroughly investigated. Thus, the first purpose of the present study is to investigate how non-native listeners' perception of Japanese length contrasts is affected by contextual factors that affect segment duration, including speaking rate, presentation con-

[a)]Portions of this work were presented in "Perception of phonemic length contrasts in Japanese by native and non-native listeners," Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, August, 2003.

[b)]Electronic mail: tajima@hosei.ac.jp

text (target words uttered in isolation versus embedded in a carrier sentence), and position within the word.

Meanwhile, numerous studies on L2 speech learning have demonstrated that, contrary to the traditional belief that adults lose neurological plasticity required for acquiring the sound system of a foreign language (Lenneberg, 1967), their production and perception abilities can be modified with experience. Both exposure to L2 (e.g., Yamada, 1995) and L2 speech training (e.g., Logan *et al.*, 1991; Lively *et al.*, 1993, 1994; Bradlow *et al.*, 1997) lead to substantial improvement in non-native listeners' ability to perceive and produce difficult L2 phonetic contrasts. In particular, Logan *et al.* (1991) and Lively *et al.* (1993, 1994) have demonstrated that "high variability perceptual training," which exposes trainees to instances of L2 phonetic categories produced in many phonetic environments and by many talkers, significantly improves listeners' ability to identify difficult L2 phonemes, such as the perception of the English /r/-/l/ contrast by Japanese listeners. Recently, this training method has been demonstrated to be effective for training prosodic properties of L2 speech as well, such as the perception of Mandarin lexical tones by English listeners (Wang *et al.*, 1999), perception of English syllables by Japanese listeners (Tajima and Erickson, 2001), and perception of Japanese length contrasts (e.g., Hirata *et al.*, 2007). However, the extent to which this training method improves perception of Japanese length contrasts has not been thoroughly investigated, especially in relation to the above-mentioned contextual factors that might affect performance. As such, the second purpose of the present study is to address the degree to which perceptual training improves English listeners' perception of Japanese length contrasts, and to assess the degree to which training generalizes to novel stimulus conditions that affect the temporal context.

A. Japanese length contrasts

Japanese can be said to have several distinct types of length contrasts that are signaled primarily by duration, depending on the type of segment involved. Phonologically, the length contrast involves the addition of an extra mora. In the present study, the following four types of length-based minimal pairs are considered (a hyphen "-" indicates mora boundaries): (1) *vowel pairs*, which contrast in the length of a vowel, e.g., /ka-do/ (corner) versus /ka-a-do/ (card); (2) *obstruent pairs*, which contrast in the length of an obstruent consonant, e.g., /ha-ke-n/ (dispatch) versus /ha-k-ke-n/ (discovery); (3) *nasal pairs*, which contrast in the length of a nasal consonant, e.g., /ta-ni-n/ (stranger) versus /ta-n-ni-n/ (person in charge); and (4) *palatal pairs*, which differ in the length of a palatal /i/-like segment, e.g., /kja-ku/ (visitor) versus /ki-ja-ku/ (statute). The first three types of contrast involve the presence or absence of a moraic vowel, moraic obstruent, or moraic nasal, respectively, signaled orthographically by distinct kana syllabary symbols. The moraic nasal is not difficult to acquire for non-native learners when they precede non-nasal consonants, e.g., /ho-n-da/ (Honda), but when it is immediately followed by another nasal consonant, its presence is signaled by the duration of the nasal segment, therefore causing potential problems for non-native learners (Uchida, 1998). The fourth contrast type, palatal pairs, is signaled orthographically by using either a subscript or full-size kana symbol that indicates the palatal sound, corresponding to the short and long members of the pair, respectively.

Partly because different orthographic conventions are used to transcribe the four contrast types in Japanese, they have often not been examined together under a common framework. However, all four contrast types can be construed as being signaled mainly by durational cues. For example, several studies have shown that the primary acoustic and perceptual cues for distinguishing short versus long vowels, obstruents, and nasals are the duration of the vowel, obstruent, or nasal segment, respectively (Fujisaka *et al.*, 1975; Uchida, 1998). Furthermore, these contrasts have been shown to be perceived in a categorical manner by native listeners (Uchida, 1998). In fact, these studies have claimed that essentially the same perceptual mechanism is employed for perceiving these contrast types. Cues for distinguishing palatal pairs have not been extensively investigated, but the primary perceptual cue for palatal pairs such as /kjaku/-/kijaku/ is likely the duration of a /i/-like vocalic interval with a high second formant (which immediately precedes another vocalic segment with a relatively low second formant, such as /a/, /o/ or /u/).

If the same perceptual mechanism is involved in perceiving the four contrast types, then improvement in the ability to perceive one type of contrast might generalize to other contrast types. On the other hand, it has been reported that some contrasts are more difficult to learn than others; for example, Toda (2003) has reported that obstruent length contrasts are more difficult to acquire than vowel length contrasts. If so, then learning to perceive one type of contrast may not be sufficient to guarantee improved ability to perceive other contrast types. The present study investigated this question by training listeners with only vowel pairs, and testing whether training generalizes to the other three contrast types, or whether the effect of training is limited to the contrast type that listeners were trained with.

B. Sources of contextual variability

The present study investigated three contextual factors that are expected to affect the perception of length contrasts by non-native listeners: speaking rate, presentation context, and within-word position. Simultaneous investigation of these factors in a single study was necessary because some of the factors were expected to interact with each other.

First, speaking rate exerts large effects on segment duration. For example, Hirata (2004a) found that the duration of a short vowel produced at a slow speaking rate is sometimes longer than that of a long vowel produced at a fast rate, and that the durational difference between short and long segments is relatively large at slow speaking rates but smaller at faster rates. These results suggest that listeners may have difficulty perceiving length contrasts if they rely on a fixed durational threshold between short and long segments, and that their perceptual performance may decline as speaking rate increases. Native Japanese listeners have been

shown to shift their perceptual boundary between phonemically short and long segments according to changes in speaking rate (e.g., Fujisaki *et al.*, 1975). However, studies conducted with non-native listeners have shown different results. For example, Toda (2003) asked listeners to identify words that belonged to continua such as /kate/-/kate:/ and /kate/-/katte/, and manipulated speaking rate by either shortening or lengthening the duration of the first vowel /a/. She found that while native listeners' perceptual boundary generally shifted according to the duration change in the preceding segment, English listeners did not show such a systematic shift. Toda's study, however, artificially manipulated the duration of segments, and did not use materials that naturally varied in speaking rate.

A recent study by Hirata *et al.* (2007) reported that two-rate training, in which native English listeners are trained using sentences produced at two speaking rates (slow and fast), is superior over one-rate training, in which listeners are trained using sentences produced at one rate (slow or fast). Listeners trained with two rates performed better than listeners trained with one rate when tested with sentences produced at various speaking rates. The present study pursued a similar question, by training listeners using words produced at a normal rate, and testing whether performance improves for words and sentences produced at slow, normal, and fast speaking rates.

Second, the presentation context, i.e., whether the target word is produced in isolation or is embedded in a carrier sentence, may also affect perception of Japanese length contrasts. For native Japanese listeners, Fujisaki *et al.* (1975) found that the perceptual boundary between phonemically short and long segments shifted as a function of speaking rate for both words uttered in isolation and words embedded in a short carrier sentence, but this adaptation was slightly stronger in sentence context than in word context, suggesting that carrier sentences provide contextual cues that facilitate judgment of speaking rate.

Whether non-native listeners would also benefit from carrier sentences, however, is unclear. Both inhibitory and facilitatory effects are conceivable. On the one hand, carrier sentences increase the amount of information that listeners need to process, and require listeners to spot the word in the sentence, while no such segmentation would be necessary if the word were presented by itself. Furthermore, words produced in sentences are typically spoken at a faster rate than the same words produced in isolation, since an increase in the number of syllables or words in a breath group typically leads to an increase in speaking rate. These factors might together make the sentence condition more difficult than the word condition. Studies on L2 phoneme perception have in fact demonstrated that non-native listeners' identification performance is poorer when the target word is in a semantically neutral carrier sentence than when it is presented in isolation (Ikuma and Akahane-Yamada, 2004). Hirata (2004b) also reported that English listeners' perception of Japanese length contrasts was worse in sentence context than word context.

On the other hand, a carrier sentence potentially provides contextual cues about overall tempo of the utterance, which could be useful in judging the phonemic length of segments. In a study examining the role of sentence context for native Japanese listeners' perception of vowel length, Hirata and Lambacher (2004) found that native Japanese listeners' perception of vowel length was poorer when the target word was excised from the carrier sentence and presented in isolation than when the target word was presented in the original carrier sentence. This suggests that the carrier sentence contained important information for accurately perceiving phonemic length. While it is not clear whether a similar disadvantage would be observed for a target word that is originally produced in isolation (as opposed to being excised from a carrier sentence), this opens the possibility that non-native listeners may also benefit from contextual cues surrounding the target word.

Finally, the position of the length contrast within the word may also affect performance. Statistical analyses of segment duration in Japanese speech databases have indicated that vowels exhibit final lengthening at the end of an isolated word or a (sentence-nonfinal) phrase, but they exhibit final *shortening* at the end of a sentence (Takeda *et al.*, 1989; Kaiki and Sagisaka, 1992). Whether within-word position affects the durational contrast between short and long segments has not been extensively investigated, but there is some limited data from acoustic measurements of vowels in isolated-word utterances (Kubozono, 2002) which indicate that the difference between short and long vowels is smaller in word-final position than nonfinal position. This suggests that perception of length contrasts might be less accurate in word-final position than other positions. Such a position effect has not been closely examined for native listeners, but studies with non-native listeners have found that errors in identifying short and long vowels were most frequent in word-final position, and significantly less frequent when the vowel appeared in a word-initial or word-medial syllable (Oguma, 2000; Minagawa-Kawai *et al.*, 2002).

To explain this position effect, it has been suggested that the effect can be attributed to the presence/absence of phonetic materials following the target segment. That is, when the target segment is word-internal, it is followed by other speech sounds that potentially provide additional timing cues that facilitate judgment of phonemic length, while no such cues are available when the segment is word-final [Minagawa-Kawai *et al.* (2002); see also Kubozono (2002)]. If this explanation holds, then one would predict that the position effect would not be observed if the word-final segment were followed by other phonetic materials, e.g., those belonging to the carrier sentence. The present study investigated this question by examining the effect of position for both words produced in isolation and words embedded in carrier sentences.

C. Perceptual training

Effects of laboratory training on English listeners' perception of Japanese length contrasts have been examined in several previous studies. For example, Yamada *et al.* (1994) trained American English listeners to identify nonwords of the form $C_1V_1C_2V_2$ where V_1 or C_2 varied in segment iden-

tity as well as phonemic length, and found that training significantly improved performance, and generalized to untrained nonwords and talkers. A series of recent studies by Hirata and her colleagues have also examined perceptual learning of Japanese length contrasts by English listeners. Hirata (2004b) investigated the effect of training using isolated words versus words embedded in carrier sentences, and found that listeners trained with words in isolation improved performance for words embedded in sentences, and vice versa. Hirata *et al.* (2007) examined the effect of training using sentences produced at two rates versus/sentences produced at only one rate, and found that both one- and two-rate training improved performance but that two-rate training yielded more robust generalization to untrained rates. Many of these studies, however, typically tested listeners with stimuli in which the length contrast occurred in fixed positions in the target word. It is therefore unclear to what extent listeners' performance varies across different positions. It is also unclear whether speaking rate and presentation context combine in a simple additive manner or whether they interact in complex ways.

Tajima *et al.* (2003b) trained native English listeners residing in Japan to identify Japanese words contrasting in phonemic length using a minimal-pair identification task. Listeners were trained with vowel pairs spoken in isolation at a normal rate, but were tested with words of various contrast types produced at three speaking rates and in two presentation contexts (in isolation or embedded in a carrier sentence). Training improved performance from 90.6% to 94.1%, but listeners' performance was very high even before training, making it difficult to assess the effectiveness of training and to evaluate the degree to which training generalizes to various conditions. One possible reason for the high accuracies in the study of Tajima *et al.* (2003b) is that the identification task had relatively low stimulus uncertainty (cf. Watson *et al.*, 1976), making the identification task fairly easy even for non-native listeners. That is, in each block of trials, speaking rate and contrast type were fixed; thus, listeners could easily predict which stimulus properties to pay attention to, and they could potentially set up a fixed perceptual criterion for judging the phonemic length of the target segment. Furthermore, stimuli were produced by professionally trained talkers, who were expected to be better able than nonprofessional talkers to produce a clear distinction between short and long segments even at multiple speaking rates. Thus, performance may have been poorer had the stimuli been produced by nonprofessional talkers.

In the present study, two experiments were carried out in order to investigate the effect of perceptual training under conditions of high contextual variability, e.g., conditions in which speaking rate, presentation context, and contrast type vary. Experiment 1 tested listeners in conditions of relatively high trial-to-trial stimulus uncertainty, and focused on how differences in contrast type and speaking rate affect performance. Rather than presenting just a single speaking rate and a single contrast type within each block of trials, as was done in the study of Tajima *et al.* (2003b), stimuli from the three speaking rates and the four contrast types were mixed and presented in a random order within the same block of trials. Such a test not only was expected to increase task difficulty, but it was also expected to be a more sensitive test of how speaking rate affected performance, and how training generalized to various untrained stimulus conditions. Experiment 2 focused on testing whether perceptual training using words produced by professionally trained talkers also improves performance on ordinary, nonprofessional talkers, who may not produce as clear a length distinction as professional talkers. Experiment 2 also focused on the effect of within-word position, and examined whether length contrasts are more difficult to identify in word-final position than other positions, and whether this position effect would be reduced if words are embedded in carrier sentences (so that word-final target segments would be followed by other phonetic materials). The experiment also examined how these positional effects are modified with perceptual training. In both experiments 1 and 2, a pretest–posttest design was employed in which a group of non-native listeners took the same test twice (dubbed test1 and test2), once before and once after training. As a control group, a different group of non-native listeners took only test1 and test2 separated by about the same number of days as the training group.

II. EXPERIMENT 1

A. Participants

Three listener groups participated. (1) Group ET (English Training): ten native Canadian English listeners who took five days of training between test1 and test2 (one male, nine females, aged 19–25, mean age=21.3). (2) Group EC (English Control): ten native Canadian English listeners who took only test1 and test2, but no training (five male, five females, aged 18–21, mean age=19.2). Listeners in groups ET and EC had no prior experience with Japanese. (3) Group JC (Japanese Control): ten native Japanese speakers (seven males, three females, aged 19–22, mean age=20.6). The English listeners participated in the experiment at Queen's University, and the Japanese listeners at ATR Laboratories. None of the listeners had any history of speech or hearing disorders.

B. Stimuli and procedure

The experiment consisted of three phases: test1, five days of training, and test2. Group ET participated in all three phases, group EC participated in test1 and test2, and group JC took the test once. Prior to the experiment, the non-native listeners were given a brief description of Japanese length contrasts, along with audio samples and English transcriptions that illustrated the four contrast types. Long vowels were transcribed as "i: e: a: o: u:" rather than "ii ee aa oo uu" because double letters such as "ee" and "oo" were likely to be misinterpreted as /i:/ and /u:/ by English listeners (a color ":" was used to mimic the IPA symbol for extra length). Long consonants were transcribed using double letters, i.e., "pp tt kk ss zz mm nn jj," or as "ssh" and "tch" for long counterparts of "sh" and "ch," respectively. The sample words were not used in the test or training.

1. Test

The test stimuli consisted of 76 real word pairs and three nonword triplets. Words in each real word pair minimally contrasted in one of the four contrast types, and were matched in word accent pattern. The word pairs were selected based on a search through a Japanese lexical database (Amano and Kondo, 2000), which contains, for over 80 000 Japanese words, subjective ratings for word familiarity, appropriateness of word accent pattern, etc. Most word pairs used in the present study had familiarity ratings of 5.0 or higher (on a scale from 1.0 to 7.0) and accent appropriateness ratings of 4.7 and higher (on a scale from 1.0 to 5.0), although nasal and palatal pairs contained greater proportions of less familiar words (due to the paucity of available minimal pairs). The word pairs were also selected so that the set as a whole was reasonably phonetically balanced. Target segments appeared in several possible positions within the word depending on the stimulus. For example, vowel length contrasts appeared in either the word-initial syllable, e.g., /kado/ versus /ka:do/ or word-final syllable, e.g., /kaze/ (wind) versus /kaze:/ (taxation), or they appeared in monosyllabic words, e.g., /ki/ (tree) versus /ki:/ (key). Nasal length contrasts appeared either in the word-initial syllable /tanin/ versus /tannin/ or in a phrase-medial position, e.g., /koi no e/ (picture of a carp) versus /koin no e/ (picture of a coin). For nasal length contrasts in word-medial position, short phrases were devised as stimuli because there were no appropriate word pairs that contrasted in nasal length in word-medial position. The three nonword triplets were of the form /ereCe/-/ere:Ce/-/ereC:e/, where C was one of the following consonants, /t s n/. The 76 real word pairs consisted of 20 vowel, obstruent, and nasal pairs each, and 16 palatal pairs. They were equally divided into four lists, each of which was to be read by a different talker. The three nonword triplets were to be read by all talkers.

Each word or nonword was read in two contexts, (1) in isolation (word context), and (2) in a sentence context in which each item was randomly embedded in one of ten carrier sentences, e.g., /ima kara __ to iimasu/ (I will say __ now). All carrier sentences had four moras preceding the target word, and either five or six moras following the target word depending on the carrier sentence. A different carrier sentence was assigned to each item across lists to be read by different talkers. The words and sentences were compiled into separate lists.

The talkers were four professionally trained native Japanese talkers (two females aged 38 and 51 and two males aged 37 and 53), who had been trained as voice actors/actresses and spoke standard Tokyo Japanese comfortably. The lists were read first at a self-selected normal rate, then at a fast rate, and finally at a slow rate. To obtain speech samples produced at sufficiently distinct speaking rates, the talkers were encouraged to utter the "fast" items at about twice the speed as the "slow" items. The recording took place in an anechoic chamber at ATR Laboratories, and was later saved as audio files at 22.05 Hz sampling frequency and 16 bit resolution.[1]

Listeners took the test in a sound-treated booth. The task was a single-stimulus, two-or three-alternative forced-choice identification task. On each trial, English transcriptions of two Japanese words comprising a minimal pair, or three nonwords comprising a triplet, appeared as clickable buttons in the computer program window. In the sentence condition, an English transcription of the carrier sentence was also presented, with the target word replaced by an underline. Simultaneously, listeners heard one of the words, or a sentence containing one of the target words, presented through headphones at a comfortable listening level. Their task was to select the word they heard by clicking the appropriate button. Listeners were able to listen to the stimulus again by clicking the "replay" button, but they were discouraged from doing so frequently. The trials were self-paced. The test consisted of 1128 trials, divided into 16 blocks of either 114 real word trials or 27 nonword trials. In each block, stimuli from each combination of the following three factors were presented: presentation context (word versus sentence), talker, and word type (word versus nonword). The order of the two presentation contexts and the four talkers was counterbalanced across listeners, but the order of the word types was fixed, such that the word trials always immediately preceded the corresponding nonword trials. Within each block of word trials, words from the four contrast types uttered at the three speaking rates were presented in a random order (38 words $\times$ 3 rates $=$ 114 trials). Within each block of nonword trials, nonwords spoken at the three rates were presented in a random order (9 nonwords $\times$ 3 rates $=$ 27 trials). Listeners were allowed to take short breaks between blocks of trials and halfway within long blocks of trials. The four test talkers were different from the talkers who appeared during training. (An unfortunate error in the experimental design resulted in some of the test words in experiment 1 appearing in both the tests and the training, contrary to the original intention which was to have no overlap between the test and training stimuli, as a genuine test of generalization of training to untrained items. Specifically, 9 of the 20 vowel pairs appeared in both the tests and the training. All other word pairs and nonwords were different from those that appeared during training. As discussed in Sec. II C, there were no consistent differences in listeners' performance between trained and untrained stimuli.)

Groups ET and EC took the same test twice, with an average of 9.1 days (7–12 days) between test1 and test2 for group ET and 9.8 days (7–14 days) for group EC. For both groups, test1 took approximately 60–90 min, while test2 took approximately 50–70 min. The JC listeners took the test only once.

2. Training

Listeners in group ET underwent five days of perceptual identification training between test1 and test2. The training stimuli consisted of 60 vowel pairs, most of which were different from the test pairs (see previous text). There were roughly an equal number of vowel pairs that contained the vowel length contrast in the initial syllable (31 pairs) and those that contained the contrast in the final syllable (27 pairs); the remaining two pairs were monosyllabic words. There were no vowel pairs that contained the contrast in word-medial position. The training words were produced in isolation at a normal speaking rate only, by five profession-

ally trained native Japanese talkers of various ages (two males and three females, aged 35–65). These talkers were different from the four talkers who appeared during the tests. These talkers were actually instructed to read a larger set of Japanese words and sentences at multiple speaking rates, similar to the test talkers, but only some of the isolated words produced at a normal rate were used for training. The recording took place in a recording studio in Tokyo, and was later saved as audio files in the same format as the test stimuli.

The training was conducted in the same laboratory environment as the tests. Three training sessions were conducted on each training day, with no more than three free days separating consecutive training days. There were 240 trials in each session, in which the 60 vowel pairs as spoken by one talker were presented two times each in a random order. Over the five training days, listeners cycled three times, in a fixed order, through the five talkers, yielding a total of 15 training sessions or 3600 training trials. Training trials were identical to the test trials except that immediate feedback was provided concerning listeners responses and that correction trials were performed, such that listeners repeated a given trial until the correct response was selected. Each training day lasted for roughly 35–60 min, with a mild tendency for sessions to become shorter as listeners accumulated training.

C. Results and discussion

In subsequent sections, listeners' performance is reported as percent-correct identification accuracies. All statistical tests in experiments 1 and 2 were conducted on arcsine-transformed values of the identification accuracies. Repeated-measures analyses of variance (ANOVA) were conducted with correction for sphericity, based on Greenhouse and Geisser's (1959) method.

An error in the experimental design resulted in 9 of the 20 vowel pairs being included in both the tests and the training, while all other items in the tests were different from those used during training, as originally intended. To examine whether group ET's performance differed between trained versus untrained test items, mean identification accuracies were computed separately for the 9 trained vowel pairs and 11 untrained vowel pairs, for each test. Results indicated that identification scores were not significantly different from each other for either test1 or test2. Thus, all subsequent analyses pooled together data from trained and untrained test items.

1. Overall performance

Figure 1 shows boxplots of the identification accuracies in test1 and test2 for groups EC and ET and accuracies in test1 for group JC. Accuracies are based on trials in all conditions of talker, rate, presentation context, and contrast type, including words and nonwords.

Figure 1 shows considerable individual variation in performance among the non-native listeners. For group ET, mean identification accuracy was 69.1% (s.d.=7.3) in test1, but rose to 76.6% (s.d.=10.3) in test2. For group EC, accuracies in test1 and test2 were 75.2% (s.d.=7.0) and 79.3% (s.d.=6.7), respectively. Mean accuracy in test1 turned out to be higher for group EC than for group ET, even though both groups were recruited from the same subject pool at the same university in Canada. Listeners in group ET improved from test1 to test2 by 7.5 percentage points on average, but group EC's mean accuracy also improved, although to a smaller degree on average (4.1 points). Group JC's mean accuracy was 98.0% (s.d.=0.8).

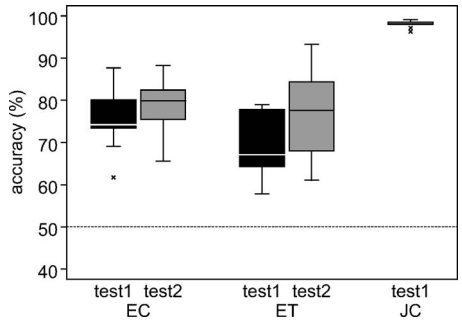

FIG. 1. Identification accuracies in experiment 1 as a function of listener group (EC, ET, JC) and test (test1, test2). The JC listeners took the test only once. Accuracies are based on trials in all conditions of talker, rate, presentation context, and contrast type, including words and nonwords. The horizontal dashed line indicates chance level performance. In this and subsequent boxplots, the horizontal line in each box indicates the median, the vertical length of the box indicates the interquartile range (the range between the lower and upper quartiles), each whisker indicates the furthest data point from the edge of the box that is not further than 1.5 times the length of the box, and individual data points indicate outliers that are further out than the extent of the whiskers.

The non-native listeners' accuracies were submitted to a two-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable and test (test1, test2) as a within-subjects variable. Results revealed a significant main effect of test [$F(1,18)=19.64$, $p<0.001$], but no significant (n.s.) main effect of group [$F(1,18)=1.49$, n.s.], or significant interaction between test and group [$F(1,18)=1.75$, n.s.]. Further pairwise comparisons of accuracies (with Bonferroni correction when family-wise error rate was set at $\alpha=0.05$) revealed that the improvement in accuracy from test1 to test2 was significant for both group ET [$t(9)=3.14$, $p<0.05$] and group EC [$t(9)=3.86$, $p<0.05$]. However, there was no significant difference in accuracy between groups ET and EC for either test1 [$t(18)=1.90$, n.s.] or test2 [$t(18)=0.57$, n.s.]. Another set of comparisons between the non-native listeners' test2 scores and native Japanese listeners' scores revealed that both group ET's test2 scores [$t(18)=6.55$, $p<0.05$] and group EC's test2 scores [$t(18)=11.93$, $p<0.05$] were significantly lower than group JC's scores.

To briefly examine how listeners' performance changed during the training sessions, mean accuracies among the ten ET listeners were computed for each of the 15 training sessions. Listeners started out at 83.7% accuracy in session 1 and ended at 91.1% accuracy in session 15, with the highest accuracy (93.5%) obtained in session 12. Generally speaking, accuracy increased across the 15 sessions, but the amount of increase was greater during the first half of training than the second half.

In summary, these results suggest that non-native listeners have considerable difficulty identifying phonemic length contrasts when they appear in Japanese words and sentences spoken at various speaking rates. Accuracies in test1 and test2 in the present study were lower than those obtained in the study of Tajima *et al.* (2003b)—87%. The primary difference between the study of Tajima *et al.* (2003b) and the present study was that speaking rate and contrast type varied from trial to trial in the latter study, while they were fixed within blocks of trials in the former. This suggests that non-native listeners have perceptual difficulties especially under conditions of considerable trial-to-trial stimulus uncertainty.

Performance during training, in which only vowel pairs produced in isolation at a normal rate were used, turned out to be relatively high, even at the beginning of training. This suggests that there was little room left for further improvement to take place during training, although accuracy did improve to some extent as training proceeded.

Perhaps because of the relative small improvement during training, group ET's improvement in accuracy from test1 to test2 was not significantly greater than the improvement observed for group EC, as indicated by the lack of a significant interaction between group and test. Thus, the results in Fig. 1 do not provide strong evidence that training *per se* led to significantly greater improvement in performance than factors such as repeated exposure to the test materials and increased familiarity with the task.

2. Contrast type

Even though group ET did not show a significantly greater increase in overall accuracy than group EC, it is possible that repeated training with vowel pairs may lead to more specific improvement in certain contrast types for group ET. To examine this, Fig. 2(a) shows the performance of groups ET and JC as a function of the four contrast types and the nonwords. Figure 2(b) shows a similar graph for group EC. Accuracies in both Figs. 2(a) and 2(b) are based on stimuli produced at all rates and presentation contexts by all talkers. Figure 2 suggest that non-native listeners' accuracies were generally lower for the nonwords than for the real words. This may be related to the fact that chance level was 33% for the nonwords but 50% for the real words (shown as dotted lines in Fig. 2). Performance appears to improve to some extent from test1 to test2 for both groups ET and EC. The most salient improvement seems to be witnessed for group ET's vowel pairs.

A three-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable and test (test1, test2) and contrast (vowel, obstruent, nasal, palatal) as within-subjects variables was conducted for the real words. If perceptual training improves performance equally on all contrast types, then one might expect to see a significant interaction between group and test, with no interactions involving contrast. On the other hand, if training improves performance only on vowel pairs or on a subset of the contrast types, then one would expect a significant three-way interaction among group, test, and contrast. Results indicated significant main effects of test [$F(1,18)=25.23$, $p<0.001$] and contrast [$F(2,31)=37.98$, $p<0.001$], and a significant

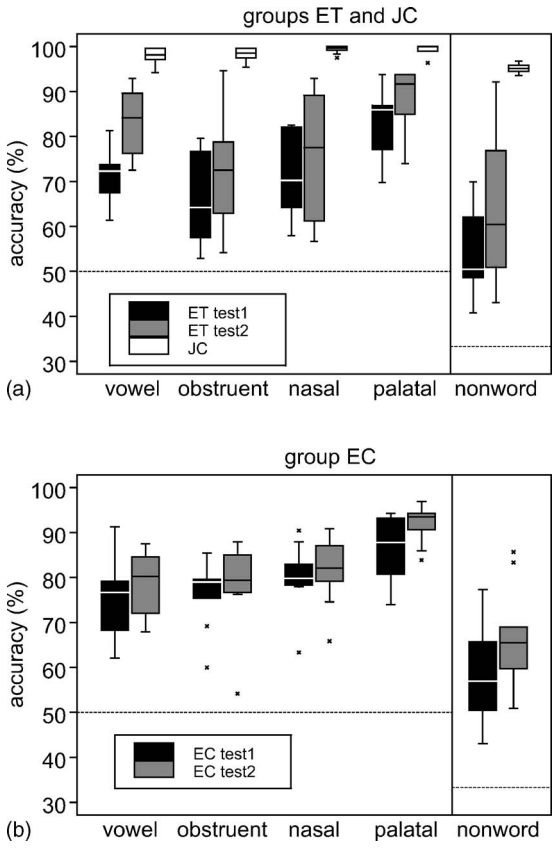

FIG. 2. Identification accuracies in experiment 1 for (a) group ET's test1 and test2 and group JC, and (b) group EC, as a function of the four contrast types and the nonwords. Accuracies are based on stimuli produced at all rates and presentation contexts by all talkers. The horizontal dashed line indicates chance level performance (50% for real words and 33% for nonwords).

test-by-contrast interaction [$F(2,38)=3.22$, $p<0.05$]. Other main effects and interactions were not significant. Further analysis of the test-by-contrast interaction with simple effect tests and multiple comparisons using Tukey's HSD revealed that for both test1 and test2, accuracy was significantly higher for palatal pairs than for the other three contrast types ($p<0.05$). For test1, accuracy for obstruent pairs was significantly lower than that for nasal pairs, but for test2, accuracy for obstruent pairs was significantly lower than that for vowel pairs ($p<0.05$). Looking at each individual contrast type, the increase in accuracy from test1 to test2 was significant for all four contrast types, although the magnitude of the increase varied across contrasts, with the greatest increase for vowel pairs (73.4% to 81.2%; $p<0.001$), followed by palatal pairs (84.6% to 90.4%; $p<0.001$), obstruent pairs (70.8% to 75.1%; $p<0.01$), and nasal pairs (76.0% to 79.2%; $p<0.05$).

In short, these results suggest that non-native listeners' ability to identify Japanese length contrasts vary depending on the contrast type involved. As for the effect of training, since there were no significant interactions involving group and test, the present results do not provide evidence that training *per se* improved performance on all contrast types, or that training improved performance on specific contrast types.

TABLE I. Mean mora duration in milliseconds (and standard deviations) for sample vowel pairs for all test talkers in experiment 1 shown for each combination of speaking rate and presentation context. Mean mora duration was computed by dividing the word duration by the number of moras in the word. Data in each cell are based on four sample vowel pairs (eight tokens). The bottom row shows the range of mean mora durations observed for tokens in each condition.

	Word			Sentence		
Talker	Slow	Normal	Fast	Slow	Normal	Fast
PF4	313.2 (20.7)	207.9 (19.8)	159.2 (13.8)	174.6 (8.4)	127.5 (7.9)	96.5 (7.0)
PF5	293.2 (34.0)	200.5 (25.1)	132.2 (21.4)	183.7 (26.0)	134.9 (22.0)	91.3 (15.4)
PM4	262.6 (23.2)	202.7 (24.7)	146.2 (15.3)	192.8 (18.7)	142.2 (8.8)	98.5 (10.7)
PM6	278.2 (29.5)	207.2 (19.3)	130.4 (9.4)	190.4 (19.1)	137.0 (10.8)	104.5 (10.9)
Mean	286.8 (32.2)	204.6 (21.5)	142.0 (19.0)	185.3 (19.5)	135.4 (14.0)	97.7 (11.9)
Range	225.4–346.1	164.7–247.7	95.2–177.5	148.5–218.3	105.2–168.7	72.7–122.9

3. Speaking rate and presentation context

To assess how well the "slow," "normal," and "fast" speaking rates were implemented by the talkers, and how speaking rate varied across talkers, word duration was measured for four sample vowel pairs for each of the four test talkers and each of the five training talkers, and mean mora duration was computed by dividing the word duration by the number of moras in the word. Table I shows results for each test talker, separately for each speaking rate and presentation context. Table II shows results for each training talker. The bottom row of each column shows the range of mean mora durations observed for the tokens measured in each condition ($N=32$ for test stimuli and $N=40$ for training stimuli).

Looking across different speaking rates and presentation contexts, Table I shows that the test talkers as a group produced three distinct rates in both word and sentence contexts. The mean mora duration across all talkers for the slow, normal, and fast rates were 286.8, 204.6, and 142.0 ms, respectively, for the word context, and 185.3, 135.4, and 97.7 ms, respectively, for the sentence context. The mean duration ratios between the fast and normal rates and between the slow and normal rates, when the normal rate is normalized to 1.00, were approximately 0.69–0.72:1.00 and 1.37–1.40:1.00, respectively. The ranges at the bottom of Table I suggest that there was considerable token-to-token variation in mean mora duration within each rate, but it appears that the professional talkers produced sufficiently distinct speaking rates. Mean mora duration was shorter overall in sentence context than in word context, reflecting a tendency for sentences to be produced at a faster speaking rate than isolated words.

TABLE II. Mean mora duration in milliseconds (and standard deviations) for sample vowel pairs for all training talkers in experiment 1. Data in each cell are based on four sample vowel pairs (eight tokens). The bottom row shows the range of mean mora durations observed for the measured tokens.

Talker	Duration (ms)	
PF1	194.8	(19.6)
PF2	161.9	(4.6)
PF3	158.0	(5.8)
PM1	189.6	(14.2)
PM3	151.1	(17.0)
Mean	171.1	(22.1)
Range	121.3–221.7	

Looking across talkers, Table I suggests that there was considerable overlap in speaking rate across the four test talkers in each condition. Differences across speaking rates and presentation contexts were much greater than differences across talkers. Table II suggests that differences in speaking rate among the five training talkers were somewhat greater than those among the test talkers. Mean mora duration ranged from 151.1 to 194.8 ms across the training talkers. It appears that three talkers (PF2, PF3, PM3) produced somewhat faster speaking rates than the other two talkers, suggesting that the training words contained some amount of speaking rate variation across talkers, even though the words were produced at a "normal rate."

Turning to the effect of speaking rate and presentation context on listeners' performance, Fig. 3(a) shows groups ET and JC's accuracies as a function of rate and context. Accuracies are based on responses to word pairs of all four contrast types. Figure 3(b) shows a similar plot for group EC. Figure 3 reveals that accuracy varied considerably depending on speaking rate. Performance was poorer at a fast speaking rate than at other rates, in both presentation contexts. Between the two presentation contexts, overall performance did not seem to be better for one context than the other. Accuracy seems to be higher overall in test2 than in test1 for both groups ET and EC. There also seems to be considerable individual variation in performance, especially for group ET.

The non-native listeners' accuracies were submitted to a four-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable, and test (test1, test2), context (word, sentence), and rate (slow, normal, fast) as within-subjects factors. If training improves perception of length contrasts in words produced at various speaking rates and presentation contexts, then one might expect to see a significant group-by-test interaction. However, if training improves performance for specific presentation contexts or speaking rates, then one would expect to see a significant interaction among group, test, and either context or rate (or both). Results indicated significant main effects of test [$F(1,18)=20.06$, $p<0.001$] and rate [$F(2,27)=95.51$, $p<0.001$], and a significant context-by-rate interaction [$F(2,32)=13.31$, $p<0.001$]. No other main effects or interactions were signifi-

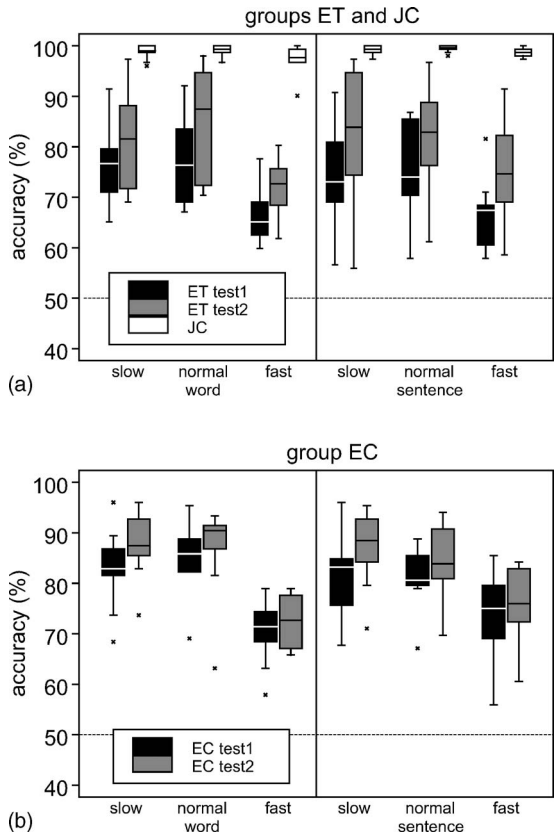

FIG. 3. Identification accuracies in experiment 1 for (a) group ET's test1 and test2 and group JC's test1, and (b) group EC, as a function of presentation context (word, sentence) and speaking rate (fast, normal, slow). Accuracies are based on responses to word pairs of all four contrast types.

cant. Further analysis of the context-by-rate interaction indicated that at the normal rate, mean accuracy was significantly higher in the word context (83.4%) than in the sentence context (80.4%; $p<0.01$), but at the fast rate, the direction reversed, such that accuracy was significantly lower in the word context (70.3%) than in the sentence context (72.8%; $p<0.05$). At the slow rate, accuracy in the word (82.1%) and sentence (81.3%) contexts did not significantly differ from each other. Looking at the effect of speaking rate for each presentation context, accuracy at the fast rate was significantly lower than accuracies at the normal and slow rates for both word and sentence contexts ($p<0.05$). Accuracies at the normal and slow rates did not significantly differ from each other.

For the the native Japanese listeners' data in Fig. 3(a), a separate two-way ANOVA with context and rate as within-subjects factors indicated that the main effect of rate was significant [$F(2,18)=10.98$, $p<0.001$]. Post-hoc tests indicated that mean accuracies at the fast rate were significantly lower than those at other rates ($p<0.05$). With the exception of one native listener who scored around 90% in the fast-rate isolated-word condition, all native listeners' scores were above 96%.

Altogether, Fig. 3 demonstrates that speaking rate has strong influences on non-native listeners' performance. Accuracies were substantially lower for stimuli produced at the fast rate than for those produced at normal or slow rates. Between the two presentation contexts, accuracy did not seem to be higher for one context than the other. Instead, presentation context was found to interact with speaking rate, such that performance for words produced in isolation was higher in the normal-rate condition but lower in the fast-rate condition than performance for words embedded in carrier sentences. Although the effect was relatively small, this suggests that presentation context and speaking rate are not independent of each other.

Between test1 and test2, accuracy was found to increase, but this was found to be the case for both groups ET and EC. Statistical tests did not indicate that the increase in accuracy was greater for group ET than for group EC, even though the increase in accuracy for group ET depicted in Fig. 3(a) seems to be somewhat greater in magnitude than those for group EC shown in Fig. 3(b). Furthermore, statistical tests failed to reveal any significant interactions involving group, test, context, and rate, suggesting that the increase in accuracy from test1 to test2 was not significantly different between word and sentence contexts, or among the three speaking rates. Thus, the present results do not provide positive evidence that training improves non-native listeners' perception of Japanese length contrasts produced at various rates and contexts.

III. EXPERIMENT 2

Results from experiment 1 suggest that overcoming variation in speaking rate and presentation context are not trivial for non-native listeners. Experiment 2 focused on a different set of factors, and investigated whether perceptual training using words produced by professionally trained talkers would also generalize to non-professional talkers' productions, in which the durational distinction between phonemically short and long segments may be less clear than productions by professional talkers. Experiment 2 also examined the extent to which performance would vary depending on the position of the length contrast within the word, and the extent to which such positional effects are modified with training.

A. Participants

All participants in experiment 2 were different from those who participated in experiment 1. Three groups of listeners participated. (1) Group ET: ten native Canadian English speakers who underwent 5 days of perceptual training between test1 and test2 (three males, seven females, aged 19–23, mean age=20.3). (2) Group EC: nine native Canadian English listeners who took only test1 and test2, but no training (five male, four females, aged 18–21, mean age =19.4). Listeners in groups ET and EC had no prior experience with Japanese. (3) Group JC: a control group of ten native Japanese speakers (six males, four females, aged 19–22, mean age=21.0). As in experiment 1, the English listeners participated in the experiment at Queen's University, and the Japanese listeners at ATR Laboratories. None of the listeners had any history of speech or hearing disorders.

B. Stimuli and procedure

The general procedure was essentially the same as experiment 1 (see Sec. II B). The stimuli were as described in the following.

The test stimuli consisted of 102 real word pairs and three nonword triplets. The real word pairs consisted of 30 vowel, obstruent, and nasal pairs each, and 12 palatal pairs, some of which were used in experiment 1. These word pairs were equally divided into six lists. Each word list was to be read by one professional talker and one nonprofessional talker, who were matched in gender and age as closely as possible. The nonword triplets were the same as those in experiment 1, and were to be read by all talkers. There were six professional talkers (three males and three females, aged 35–66); five of the six talkers appeared in experiment 1 as training talkers. There were also six nonprofessional talkers (three males and three females, aged 35–65), who spoke standard Tokyo Japanese comfortably but had never received special training as voice actors/actresses. Each talker read the words and nonwords in two contexts (in isolation and in one of ten carrier sentences as in experiment 1) at a self-selected normal speaking rate. The professional talkers also read the lists at slow and fast rates, but only normal-rate stimuli were used as test stimuli. Recording of the professional talkers took place in a recording studio in Tokyo. For nonprofessional talkers, recording was made only at a normal rate because they were likely to have difficulty reliably producing length contrasts at different rates without some practice. The recording of the nonprofessional talkers took place in an anechoic chamber at ATR Laboratories.

The test consisted of 1032 trials, divided into 24 blocks of either 34 real word trials or nine nonword trials. In each block, stimuli from each combination of the following factors were presented: presentation context (word, sentence), talker, and word type (word, nonword). The order of the two presentation contexts was counterbalanced across listeners, but the word trials always immediately preceded the corresponding nonword trials. The 12 talkers were pseudorandomly ordered in such a way that no more than two talkers from the same group (professional or nonprofessional) appeared consecutively. Within each block of word trials, 34 words from the four contrast types were presented in a random order. Within each block of nonword trials, the nine nonwords were presented in a random order. None of the 12 test talkers appeared during training as training talkers.

The ET and EC listeners took the same test twice, with an average of 8.8 days (8–11 days) between test1 and test2 for group ET and 9.1 days (7–15 days) for group EC. The JC listeners took the test once.

The training stimuli were the same 60 vowel pairs as those used in experiment 1. As mentioned before, 31 of the 60 pairs contained the vowel length contrast in the initial syllable, 27 pairs contained the contrast in the final syllable, and 2 pairs were monosyllabic words. The 60 word pairs were all different from the test words. The words were produced in isolation at a normal rate, by a different set of professionally trained talkers (three females and two males, aged 27–53) from the training talkers in experiment 1. Some of the training talkers in experiment 2 appeared as test talkers in experiment 1. None of the training talkers appeared as test talkers in the present experiment.

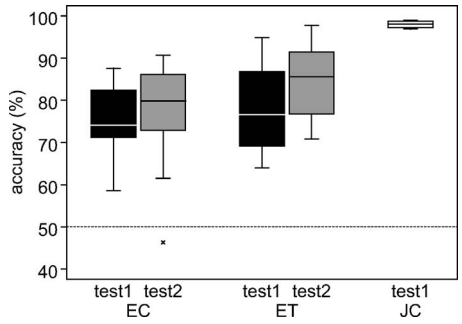

FIG. 4. Identification accuracies in experiment 2 as a function of listener group (EC, ET, JC) and test (test1, test2). The JC listeners took the test only once. Accuracies are based on trials in all conditions of talker, presentation context, and contrast type, including words and nonwords.

C. Results and discussion

1. Overall performance

Figure 4 shows the identification accuracies in test1 and test2 for groups EC and ET and accuracies in test1 for group JC. Accuracies are based on trials in all conditions of talker, presentation context, and contrast type, including words and nonwords.

For group ET, mean identification accuracy was 78.5% (s.d.=10.4) in test1, but rose to 84.4% (s.d.=9.0) in test2. For group EC, accuracies in test1 and test2 were 74.2% (s.d.=10.1) and 75.9% (s.d.=14.3), respectively. Group JC's mean accuracy was 98.0% (s.d.=0.7).

A close look at Fig. 4 reveals that one listener in group EC (EC01) scored substantially lower in test2 (46.3%) than other listeners in the same condition. This score was more than 2 s.d. lower than the mean score in test2. This listener scored 58.6% in test1, resulting in a 12.3-point decrease in accuracy from test1 to test2. Since inclusion of EC01's data effectively lowers group EC's mean scores in test2, excluding his data results in a more conservative test of the experimental hypotheses. If this listener's data are excluded, then group EC's accuracies in test1 and test2 were 76.1% (s.d. =8.8) and 79.6% (s.d.=9.7), respectively.

The non-native listeners' data in Fig. 1 were submitted to two kinds of tests, an ANOVA with EC01's data included, and an ANOVA without EC01's data. Both tests were two-way repeated-measures ANOVAs with group (ET, EC) as a between-subjects variable and test (test1, test2) as a within-subjects variable. When EC01's data were included in the analysis, results revealed a significant main effect of test $[F(1,17)=17.75, p<0.001]$, and a significant interaction between group and test $[F(1,17)=4.69, p<0.05]$. However, when EC01's data were excluded, results revealed a significant main effect of test $[F(1,16)=43.25, p<0.001]$ but no significant interaction between group and test $[F(1,17)=3.54$, n.s.]. Inclusion versus exclusion of EC01's data did not lead to different results for the analysis of talker type, contrast type, and position (Secs. III C 2, III C 3, and III C 4).

TABLE III. Mean identification accuracies in experiment 2 for each professional (pro) and nonprofessional (nonpro) talker (F=female, M=male), for group ET's test1 and test2 and for group JC. Accuracies are based on trials in all combinations of contrast type and presentation context, including words and nonwords.

Talker type	Talker	ET test1	ET test2	JC
Pro	PF1	76.6	84.2	97.9
Pro	PF2	80.0	83.5	98.5
Pro	PF3	78.4	85.3	98.4
Pro	PM1	82.9	84.1	99.1
Pro	PM2	80.5	85.9	98.5
Pro	PM3	79.8	86.6	98.1
Mean		79.7	84.9	98.4
Nonpro	NF1	80.6	86.2	98.8
Nonpro	NF2	77.8	84.1	97.8
Nonpro	NF3	79.3	85.5	97.8
Nonpro	NM1	78.0	85.3	97.8
Nonpro	NM2	77.1	82.1	93.4
Nonpro	NM3	69.7	74.8	91.3
Mean		77.1	83.0	96.1

When group ET's performance during training was briefly examined, it was found that listeners started out at 90.6% accuracy in session 1 and ended at 97.5% accuracy in session 15, with the highest accuracy (98.3%) obtained in session 13. As in experiment 1, accuracy more or less increased across the 15 sessions, but the amount of increase was greater during the first half of training than the second half.

In short, these data provide mixed results concerning whether perceptual training significantly improves non-native listeners' overall identification performance. The results varied depending on whether certain data that appear as outliers are included in the analysis or not. Even if there were an effect of training, the magnitude of the effect appears to be small (an increase of 5.9 percentage points). This may be related to the fact that listeners' performance during training was relatively high even at the beginning, leaving little room for further improvement to take place. Thus, it appears that the perceptual training method in the present study did not lead to a substantial improvement in overall performance. The following sections examine whether training modifies non-native listeners' perceptual tendencies in more subtle ways.

2. Talker type

To examine how accuracy varied across talkers and talker types, Table III shows identification accuracies separately for each talker, for groups ET and JC. Accuracies are based on trials in all combinations of contrast type and presentation context, including words and nonwords. Comparison of the two talker types indicates that mean accuracies were slightly lower for the nonprofessional talkers than for the professional talkers. A closer look at the individual talkers' accuracies reveals that among the six nonprofessional talkers, talker NM3 showed the lowest accuracies in all three tests (ET test1, ET test2, and JC). Talker NM2 also showed somewhat lower accuracies than other nonprofessional talkers in group ET's test2 and the Japanese listeners' test. The remaining nonprofessional talkers (NF1–NF3 and NM1) showed accuracies that were comparable to those of the professional talkers. A two-way ANOVA was carried out on the individual talkers' data in Table III, with talker type (professional, nonprofessional) as a between-subjects factor and test (ET test1, ET test2, JC) as a within-subjects factor. While there was a significant main effect of test [$F(2,20)=732.45$, $p<0.001$], there was no significant main effect of talker type [$F(1,10)=2.21$, n.s.] or a significant interaction [$F(2,20)=0.21$, n.s.], suggesting that there were no significant differences in listeners' performance between the professional and nonprofessional talkers. A similar analysis for group EC (not shown) also showed no significant differences in accuracy between the two talker types.

Recall that training stimuli in the present study were all produced by professionally trained talkers. The absence of a significant difference in overall accuracy between the two talker types or a significant interaction between talker type and test suggests that perceptual training using only productions by professional talkers does not have unequal effects on non-native listeners' ability to perceive productions by professional versus ordinary, nonprofessional talkers.

3. Contrast type

To test again for whether or not training has differential effects on non-native listeners' perception of trained versus untrained contrast types, Fig. 5(a) shows group ET's accuracies in test1 and test2, as well as group JC's accuracies, as a function of the four contrast types and the nonwords. Figure 5(b) shows a similar graph for group EC. Accuracies are based on stimuli produced by all talkers and in both presentation contexts. Listener EC01's data are omitted from Fig. 5(b) and the statistical analysis.

Figure 5 suggests that non-native listeners' accuracies were lowest for nonwords, and higher for the palatal contrasts than for other real word pairs. Also, the greatest increase in accuracy from test1 to test2 seems to be observed for group ET's vowel pairs. These patterns are similar to Fig. 2 in experiment 1.

A three-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable and test (test1, test2) and contrast (vowel, obstruent, nasal, palatal) as within-subjects variables was conducted for the real words. Listener EC01's data were excluded.[2] Results revealed significant main effects of test [$F(1.17)=36.07$, $p<0.001$] and contrast [$F(2,26)=12.01$, $p<0.001$], and a significant three-way interaction among group, test, and contrast [$F(2,33)=4.61$, $p<0.05$]. Further analysis of the three-way interaction was carried out by examining the group-by-test interaction for each contrast type. For vowel pairs, the test-by-group interaction was highly significant ($p<0.001$). Further examination of this interaction revealed that the increase in accuracy from test1 to test2 was significant for group ET (78.6% to 89.2%; $p<0.001$) but not for group EC (75.1% to 77.3%). For obstruent, nasal, and palatal pairs, the group-by-test interactions were not significant, suggesting that increases in accuracy from test1 to test2 did not significantly differ between group ET (obstruent: 80.5% to 83.3%; nasal: 81.8% to

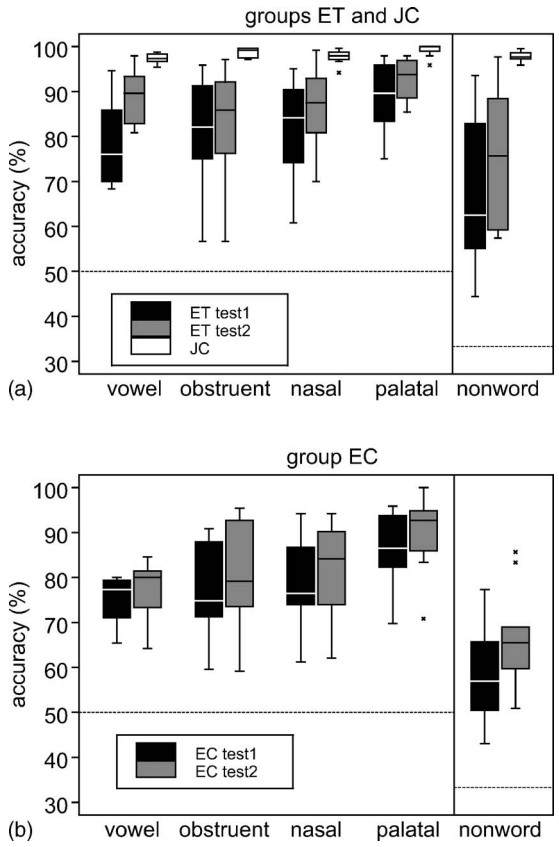

FIG. 5. Identification accuracies in experiment 2 for (a) group ET's test1 and test2 and group JC, and (b) group EC, as a function of the four contrast types and the nonwords. Accuracies are based on stimuli produced by all talkers and in both presentation contexts. The horizontal dashed line indicates chance level performance (50% for real words and 33% for nonwords).

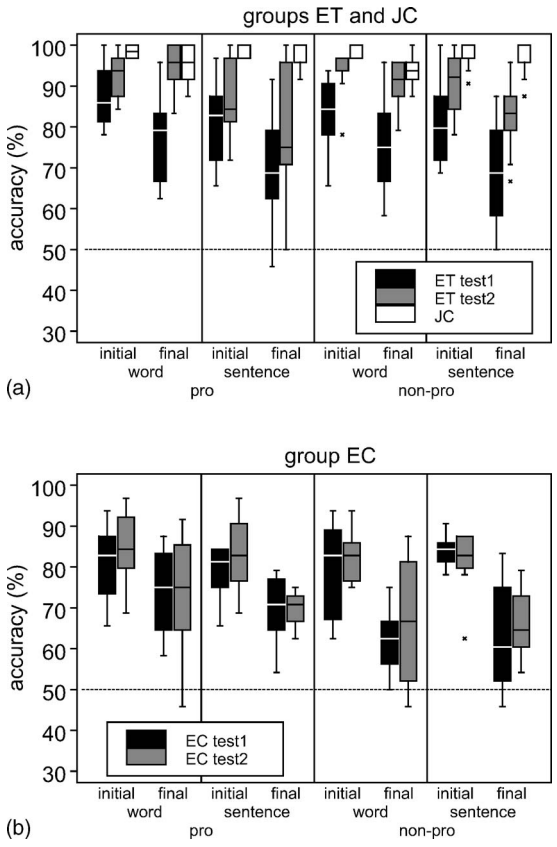

FIG. 6. Identification accuracies in experiment 2 for (a) group ET's test1 and test2 and group JC, and (b) group EC, as a function of within-word position (initial, final) separately for professional and nonprofessional talkers and for word and sentence contexts. Accuracies are based on polysyllabic vowel pairs produced by all test talkers.

86.0%; palatal: 88.5% to 92.8%) and EC (obstruent: 77.3% to 80.7%; nasal: 78.7% to 81.6%; palatal: 86.3% to 89.7%).

Put together, these results suggest, first, that non-native listeners show varying degrees of difficulty with different types of length contrasts, consistent with the results in experiment 1. Second, a significant group-by-test interaction for vowel pairs suggests that training significantly improved non-native listeners' perception of contrast types that listeners were specifically trained with, even though the specific word pairs used in the tests were different from those presented during training. Finally, the lack of significant group-by-test interactions for obstruent, nasal, and palatal pairs suggests that training did not reliably generalize to contrast types that listeners were not trained with.

4. Within-word position

To examine how position of the length contrast within the target word affected listeners' performance, Fig. 6(a) shows identification accuracies for groups ET and JC as a function of the position of the length contrast within the target word, separately for professional and nonprofessional talkers and for word and sentence contexts. Accuracies are based on polysyllabic vowel pairs produced by all test talkers. Figure 6(b) shows a similar plot for group EC. Listener EC01's data are omitted from Fig. 6 and the statistical analyses. Figure 6 shows a general tendency for non-native listeners' accuracies to be higher for length contrasts appearing in word-initial syllables than those in word-final syllables. This seems to be evident for group EC's test1 as well as test2. For group ET, this tendency appears to be somewhat weaker in test2 than in test1.

The non-native listeners' data were submitted to a five-factor repeated-measures ANOVA with listener group (ET, EC) as a between-subjects variable, and with test (test1, test2), talker type (professional, nonprofessional), context (word, sentence), and position (initial, final) as within-subjects factors. If perceptual training improves performance for specific talker types, contexts, or positions, then listener group and test should show significant interactions with these factors.

Results revealed that the main effects for all five factors were significant: listener group [ET: 83.2%; EC: 75.0%; $F(1,16)=6.66$, $p<0.05$], test [test1: 76.1%; test2: 83.0%; $F(1,16)=61.56$, $p<0.001$], talker type [pro: 80.5%; nonpro: 78.6%; $F(1,16)=10.00$, $p<0.01$], context [word: 81.7%; sentence: 77.4%; $F(1,16)=16.37$, $p<0.001$], and position [initial: 84.7%; final: 74.4%; $F(1,16)=51.05$, $p<0.001$]. These results suggest that non-native listeners' performance was strongly affected by the position of the length contrast within the word; performance was poorer by roughly 10 percentage points on average when the length contrast appeared in the word-final syllable than when it appeared in the word-initial syllable. Talker type also significantly affected perfor-

mance, but the size of the effect (a 1.9 percentage point difference) was very small, and Sec. III C 2 clearly shows that this was primarily due to low accuracies associated with one or two of the nonprofessional talkers. Context was also found to affect performance; accuracies were slightly higher when the target word appeared in isolation than when it was embedded in a carrier sentence. This perhaps suggests that carrier sentences have inhibitory, rather than facilitatory, effects on non-native listeners' performance.

In addition to the main effects, the following interactions were significant: listener group by context, talker type by context, listener group by test, listener group by position, talker type by position, and listener group by test by talker type by position. Analysis of the listener-group-by-context interaction indicated that group ET's accuracies were significantly higher in the word context (86.7%) than in the sentence context (79.8%) ($p<0.001$), while group EC's accuracies in the word (75.4%) and sentence (74.5%) contexts did not significantly differ from each other. Next, analysis of the talker-type-by-context interaction indicated that for words produced in isolation, accuracies were significantly higher for professional talkers' productions (83.7%) than for nonprofessional talkers' productions (79.7%; $p<0.001$), but for words embedded in carrier sentences, accuracies did not significantly differ between professional (77.4%) and nonprofessional (77.4%) talkers. This result suggests that professional talkers may have produced isolated words with clearer perceptual cues for length contrasts than nonprofessional talkers.

Finally, the remaining four interactions all involved some or all of the four factors: listener group, test, talker type, and position. Thus, analysis of the highest-order, four-way interaction is reported here. This was done by examining the interaction among listeners group, test, and position for each level of talker type. For professional talkers, group ET showed significant increases in accuracy from test1 to test2 in both the initial position (84.2% to 89.5%; $p<0.05$) and the final position (73.8% to 86.5%; $p<0.001$), with a much larger increase in the final than initial position. Group EC, on the other hand, did not show significant increases in accuracy in either the initial (79.9% to 84.0%) or the final (71.9% to 71.6%) position. For nonprofessional talkers, group ET showed about the same level of significant increase in accuracy in the initial and final positions (76.7% to 89.2%; $p<0.001$), but group EC again did not show significant increase in accuracy (72.1% to 74.1%). These results suggest that training significantly improved non-native listeners' performance, for length contrasts appearing in both word-initial and word-final syllables. Training significantly improved accuracy not just for professional talkers' productions but also for nonprofessional talkers' productions. The largest improvement was observed for word-final length contrasts produced by professional talkers. In fact, even though accuracies were consistently lower for word-final contrasts than word-initial contrasts in most conditions even after training, group ET's test2 accuracies showed no significant differences between the initial (89.5%) and final (86.5%) positions for professional talkers' productions.

Group JC's data were submitted to a three-way ANOVA with talker type, context, and position as within-subjects factors. Results revealed a significant main effect of position [$F(1,9)=15.58$, $p<0.01$] and a significant context-by-position interaction [$F(1,9)=5.25$, $p<0.05$]. Further analysis of the context-by-position interaction using simple effects test revealed that accuracy was significantly lower in final position than initial position for isolated-word stimuli [$F(1,9)=18.72$, $p<0.001$].

In short, non-native listeners' performance was generally much poorer when the length contrast appeared in word-final syllables than when they appeared in word-initial syllables. Perceptual training, however, significantly improved performance for both word-initial and word-final length contrasts. Training using professional talkers' productions also significantly generalized to nonprofessional talkers' productions.

IV. GENERAL DISCUSSION

The present study assessed the extent to which adult non-native listeners' perception of Japanese length contrasts can be modified with identification training, and the extent to which factors such as contrast type, speaking rate, presentation context, within-word position, and talker type affected performance before and after training.

A. Overall effect of perceptual training

In both experiments 1 and 2, both the trained listeners as well as the untrained control listeners improved performance from test1 to test2. Even though group means showed greater improvement in accuracy for trained listeners than untrained listeners, statistical tests failed to show significant differences in the amount of improvement between trained and untrained listeners. Thus, results from the present study do not provide strong evidence that perceptual identification training improves non-native listeners' overall ability to identify Japanese length contrasts. These results are not in line with past studies that have used the same training paradigm to improve L2 learners' perception of L2 segmental contrasts (e.g., Logan *et al.*, 1991; Bradlow *et al.*, 1997), L2 tones (Wang *et al.*, 1999), and L2 syllables (Tajima and Erickson, 2001). However, these results seem to echo the finding of Hirata *et al.* (2007) that training improved English listeners' perception of Japanese vowel length contrasts only to a small extent (9.1 percentage points).

Several explanations are possible for this outcome. First, the tests in the present study may have been excessively long; each test in experiments 1 and 2 consisted of 1128 and 1032 trials, respectively. These are much greater than the number of test trials employed in other training studies, e.g., 32 trials (16 pairs) in the study of Logan *et al.* (1991), 100 trials in the study of Wang *et al.* (1999), and 180 trials in the study of Hirata *et al.* (2007). A large number of trials was necessary in the present study because listeners were to be tested with various combinations of stimulus properties. However, repeated exposure to stimuli in various conditions and increased familiarity with the task may have led to some

improvement during the tests, even in the absence of explicit feedback. This may partly account for group EC's improvement in accuracy between test1 and test2.

Second, the perceptual training employed in the present study may not have been set at an appropriate level of difficulty. Listeners' identification performance during the training sessions started out at a relatively high level in both experiment 1 (83.7%) and experiment 2 (90.6%), leaving little room for performance to improve during training.

Finally, in connection with the previous point, the training stimuli may not have contained sufficient stimulus variability to lead to robust perceptual improvement. The training stimuli in the present study were all vowel pairs produced in isolation at a normal rate. Only stimuli in this limited set of conditions were used during training so as to test for generalization to untrained conditions, and to assess the amount of variability necessary for achieving robust training effects. Past studies have shown that high stimulus variability facilitates the formation of new L2 phonetic categories (e.g., Logan *et al.*, 1991; Lively *et al.*, 1993, 1994; Bradlow *et al.*, 1997). It appears that normal-rate, isolated-word training using vowel pairs only is not sufficient to yield robust training effects.

Even though perceptual training did not lead to significant improvement in overall performance, training did seem to have subtle effects on performance, improving listeners' accuracies in some conditions but not others, as discussed in the following.

B. Effect of contrast types

Results from both experiments 1 and 2 indicated that, among the four contrast types examined (vowel, obstruent, nasal, palatal), palatal pairs had the highest accuracies, while the other three contrast types did not show consistent relative rankings. One potential reason for the high accuracies among palatal pairs is that some of the pairs used in the present study, although construed in Japanese phonology as phonemic length contrasts, can be regarded as involving the presence versus absence of the palatal /i/-like segment characteristic of this contrast, rather than involving a durational contrast; for example, the pair "shaku" (serving saké) and "shiyaku" (reagent) can be seen to differ by whether the palatal portion (underlined) is absent or present, while the pair "kyaku" and "kiyaku" differs by the relative duration of the palatal portion. Non-native listeners may therefore have been able to identify some of these pairs based on the presence versus absence of palatal segments rather than their duration.

As for the effect of training, both Fig. 2(a) from experiment 1 and Fig. 5(a) from experiment 2 suggest that the increase in accuracy from test1 to test2 was greater for group ET's vowel pairs than for other word pairs in group ET or for group EC. However, results of statistical tests from the two experiments diverged, with a significant group-by-test-by-contrast interaction found in experiment 2 but not in experiment 1. It is not clear why divergent results were obtained between the two experiments. One possible reason may be that the variation in speaking rate in experiment 1 may have reduced the effect of training, even for vowel length contrasts which listeners were trained with during training. Since all the stimuli in experiment 2 were produced at the normal rate, listeners may have been able to benefit more from training, especially for contrast types that they were trained with.

Despite the divergent statistical results, training does seem to improve accuracy for length contrasts that listeners were trained with. That is, for group ET, accuracy for vowel pairs increased from 71.4% to 83.4% in experiment 1 and from 78.6% to 89.2% in experiment 2, while for group EC, accuracy only rose from 75.3% to 79.0% in experiment 1 and from 73.3% to 74.1% in experiment 2. Given that most of the test words were different from the training words, this suggests that training generalized to untrained words of the same contrast type.

As for whether training generalized to untrained contrasts, the present data do not provide strong evidence that it does. Data from both experiments suggest that the increases in accuracy from test1 to test2 observed for group ET are not greater than those observed for group EC. These results therefore suggest that the effect of training may be restricted to the specific contrast type that listeners are trained with.

The lack of significant generalization of training to untrained contrast types seems to have important theoretical and practical implications. From a theoretical standpoint, several studies have claimed that essentially the same perceptual mechanisms are used in perceiving various types of length contrasts, in the sense that segment duration serves as the primary perceptual cue for phonemic length (Fujisaki *et al.*, 1975; Uchida, 1998). Under this view, training would be expected to lead to similar levels of improvement in performance for trained as well as untrained contrast types. However, the lack of generalization may suggest that there may be fundamental differences among the contrast types. For example, it has been reported that perceptual sensitivity to durational modifications in segment duration varies depending on the type of speech sound involved, such that sensitivity is higher for vowels than consonants (Huggins, 1972; Kato *et al.*, 2002). If so, then training using vowel pairs, which may be relatively easy to perceive, may not yield improvement in other contrast types, which may be relatively difficult. Furthermore, the four contrast types differ in the phonetic environment in which they appear. For example, vowel length contrasts are typically preceded or followed by consonants, and they form the nucleus of syllables, while obstruent and nasal length contrasts are preceded and followed by vowels, and do not form the nucleus of syllables. Such structural differences may account for why vowel length training does not straightforwardly transfer to consonant length contrasts.

From a practical standpoint, lack of generalization suggests that training listeners with just one type of length contrast does not guarantee improved perception of other contrast types. Training involving multiple contrast types might be necessary to improve non-native listeners' perception of various contrast types. Further research is necessary to determine the extent to which perception of the various contrast types are independent of one another.

C. Effect of speaking rate and presentation context

Non-native listeners' identification accuracies were found to be affected by the speaking rate and presentation context of the test stimuli. In both the word and sentence contexts, performance was lowest for the fast rate (70.3% on average), and higher for the slow (78.6%) and normal (79.6%) rates. The low accuracy at the fast rate likely stems from the fact that the durational difference between phonemically short and long segments tends to be smallest at a fast rate (Hirata, 2004a; Hirata and Whitonm, 2005) making this condition more difficult for non-native listeners than other rate conditions.

Following this line of reasoning, one would expect that non-native listeners' accuracy would be higher for the slow rate than for the normal rate, since the phonemic length distinction tends to be most salient at a slow rate. No systematic differences in accuracy, however, were found between the slow and normal rates in experiment 1. One possible reason for this is that listeners may tend to respond to the stimuli based on an "average" speaking rate among all the stimuli presented in the test. Because speaking rate varied from trial to trial in experiment 1, listeners may not have been able to fully compensate for the rate variation, and may to some extent have performed the task based on a perceptual criterion that applies to tokens produced at an average rate that lies somewhere in the middle of the range of speaking rates encountered. To the extent that this strategy is adopted, this would yield better performance for normal-rate stimuli (which are close to the average rate) and would tend to reduce performance for slow-rate and fast-rate stimuli (which both diverge from the average rate). There is some evidence from previous work that non-native listeners' accuracies were sometimes higher for normal-rate stimuli than slow-rate or fast-rate stimuli (Tajima *et al.*, 2003a). If this interpretation is correct, then accuracies for the slow-rate condition in the present study may not have been as high as expected because the salience of perceptual cues may have been be canceled out by this average rate effect.

Embedding the target word in carrier sentences did not lead to consistently higher or lower accuracies than presenting the word in isolation, but presentation context was found to interact with speaking rate, such that accuracies in the word context were higher than in the sentence context at the normal rate, but *lower* at the fast rate. Such an interaction points to the importance of examining the two factors at the same time. The source of this interaction is not entirely clear. One possibility, although speculative, is that carrier sentences may have facilitatory effects under conditions in which the target word itself contains relatively weak perceptual cues for phonemic length, while they may have inhibitory effects in other conditions. That is, when the target word itself contains relatively weak phonetic cues for phonemic length, as in the case for fast-rate stimuli, non-native listeners may benefit from contextual cues provided by the carrier sentence. On the other hand, when the target word contains sufficiently salient perceptual cues, as might be the case for normal- and slow-rate stimuli, non-native listeners may not need to rely on contextual cues provided by carrier sentences. Instead, carrier sentences may impose additional processing demands on non-native listeners, thus hindering performance rather helping it (e.g., Ikuma and Akahane-Yamada, 2004).

It is worth noting that native Japanese listeners' accuracies were very high at all speaking rates and presentation contexts. This was so even under conditions in which target words, whose speaking rate varied from trial to trial, were presented in isolation with no other contextual information. This suggests that the stimuli contained sufficient perceptual cues for identifying the length contrasts, and that native Japanese listeners, but not native English listeners, were able to utilize those cues to identify phonemic length.

As for the effect of training, statistical tests in experiment 1 indicated that group and test did not significantly interact with context or rate. This suggests that the effects of context and rate mentioned earlier applied equally to both listener groups and to both test1 and test2, with no systematic differences between groups or tests. Thus, the present findings do not provide evidence that training improves listeners' ability to cope with variation in speaking rate and presentation context. Even though speaking rate varied to some extent across the five training talkers in experiment 1 (mean mora duration of 151–194 ms according to Table II), the variability was much greater for the test stimuli (mean mora duration of 98–287 ms according to Table I), which were produced at three speaking rates and presented in a random order across trials. The relatively small variability in the training stimuli may not have been sufficient to modify non-native listeners' perceptual strategies. One way to improve the effectiveness of training may be to increase the variability in speaking rate during training. In fact, Hirata *et al.* (2007) have recently reported that training non-native listeners with sentences produced at two rates (e.g., slow and fast) leads to a more robust training effect than does training with sentences produced at only a single rate (e.g., slow only or fast only). One question that remains open for future research is whether multiple-rate training is equally effective with isolated words as it is with words embedded in sentences.

D. Effect of talker type and within-word position

Professionally trained talkers were recruited in experiment 1 since they were expected to be better able than nonprofessional talkers to produce speech at distinct speaking rates while maintaining clear distinctions between phonemically short and long segments. Experiment 2 tested whether there were in fact differences in identification accuracy between professional and nonprofessional talkers' productions, and whether perceptual training using professional talkers' productions would generalize to nonprofessional talkers' utterances. Results from experiment 2 indicated that there were no significant overall differences in accuracy between the two talker types. However, some subtle differences were observed. For example, for words produced in isolation, mean accuracies were significantly higher for professional talkers' productions (83.7%) than for nonprofessional talkers' productions (79.7%; see Sec. III C 4), suggesting that professional talkers produced clearer perceptual cues for phonemic

length in isolated-word productions than nonprofessional talkers. This gives some support for the original motivation to use professional talkers' productions during training. When the effect of training was examined, group ET showed significantly greater improvement than group EC for both professional and nonprofessional talkers' productions. Thus, training significantly generalized to utterances produced by ordinary, nonprofessional talkers.

Results from experiment 2 suggest that the position of the length contrast within the target word had a very strong effect on non-native listeners' performance. Even native Japanese listeners' performance was somewhat poorer for word-final contrasts in the word context compared to other conditions. The effect of position found in the present study is in agreement with past studies that also found poorer performance in word-final position than initial position (Oguma, 2000; Minagawa-Kawai *et al.*, 2002).

One explanation offered by previous studies (Oguma, 2000; Minagawa-Kawai *et al.*, 2002) for the lower accuracy in word-final position than word-initial position was the absence of phonetic materials in final position. This was predicted to make judgment of segment duration relatively difficult in word-final position (cf. Kubozono, 2002) However, contrary to the prediction, accuracy was relatively low in final position regardless of presentation context; no significant interaction between context and position was found in experiment 2. Since the target word in the present study was always followed by the voiceless stop /t/ of the particle /to/ when produced in carrier sentences, the target word-final segment was likely to be immediately followed by acoustically and perceptually salient speech events. Furthermore, since the particle /to/ usually forms a prosodic word with the preceding word, durational variability between the target word and the particle should be no greater than that observed within the target word (cf. Warner and Arai, 2001). If so, then the availability of temporal cues for phonemic length could be considered to be comparable between final and nonfinal positions. The difference in performance between word-initial and final positions therefore cannot be simply attributed to the presence versus absence of following phonetic materials.

An alternative explanation for the difference in accuracy between word-initial and word-final positions is the presence of pitch-related cues in addition to durational cues for length contrasts in this position. Words in Tokyo Japanese typically have either a low-high or high-low tone pattern in the first two moras, resulting in a rising or falling fundamental frequency contour. For word pairs that contain a length contrast in initial position, the contrast occurs entirely or partially within the first two moras of the word, thus causing the length contrast to be associated with different tone patterns, e.g., */kạ́-dò/* versus */kạ́-ạ̀-do/* (target segments are underlined, mora boundaries are indicated with a hypen "-," and high and low tones in the first two moras are marked with /◌́/ and /◌̀/ diacritics, respectively). This means that short and long vowels that occur in word-initial position are often associated with systematically different fundamental frequency contours in addition to differences in duration, while length contrasts appearing in other positions are typically not associated with such tone differences. The availability of such secondary cues may have made length contrasts easier to identify in initial position than other positions. Pitch-related cues have been shown to serve as secondary cues to phonemic length for native Japanese listeners (Omuro-Hayashida, 1999; Kinoshita *et al.*, 2002). Non-native listeners have been reported to rely more on durational cues than pitch cues (Tabuchi *et al.*, 1997; Omuro-Hayashida, 1999). However, the degree to which the English listeners' high accuracies for word-initial length contrasts can be attributed to pitch-related cues remains unclear.

As for the effect of training, group ET in experiment 2, which underwent training, significantly improved performance for both word-initial and word-final length contrasts, for both professional and nonprofessional talkers' productions. In contrast, group EC, which did not undergo training, did not significantly improve performance in any of the conditions. This demonstrates that perceptual training was effective at improving non-native listeners' perception of Japanese vowel length contrasts.

A closer look at the pattern of improvement for group ET indicated that the level of improvement was comparable between initial and final positions (from 76.7% to 89.2% across the two positions) for nonprofessional talkers' productions, but was smaller for word-initial contrasts (from 84.2% to 89.5%) than word-final contrasts (from 73.6% to 86.5%) for professional talkers' productions. Since accuracy for word-initial contrasts for professional talkers' productions was already relatively high in test1 (84.2%), accuracy in this condition may not have increased as much as in other conditions due to ceiling effects. Aside from this difference, levels of improvement from test1 to test2 were comparable between word-initial and word-final contrasts, and between professional and nonprofessional talkers' productions. This again supports the notion that training significantly generalized to nonprofessional talkers' utterances.

Furthermore, results did not reveal significant interactions involving listener group, test, and context. This provides some suggestion that even though listeners were trained using words in isolation, the improvement in performance for words embedded in carrier sentences (from 75.1% to 84.4% on average) was not significantly smaller than that for words produced in isolation (from 80.6% to 92.8%), although the former was slightly smaller than the latter. Concerning generalization of training to untrained contexts, Hirata (2004b) has found that listeners who were trained using words in isolation did not improve performance on words embedded in sentences as well as on words in isolation. Additional research is needed to determine the extent to which training generalizes to untrained contexts.

In conclusion, non-native listeners who were trained to identify Japanese vowel length contrasts did not show greater overall improvement in performance compared to control listeners who did not receive training. However, training seems to affect performance in more subtle ways, modifying performance in some conditions but not in others. Specifically, training improves perception of contrast types that listeners are trained with, generalizes to productions by nonprofessional talkers, and improves perception of length

contrasts that occurs in positions in the word that are originally difficult. However, training does not generalize to contrast types that listeners are not trained with, nor does it significantly improve perception of words and sentences produced at various speaking rates. Further research is needed to clarify ways to refine the training methods so as to yield more robust training effects.

ACKNOWLEDGMENTS

We are grateful to Paul Iverson, Yukari Hirata, and two anonymous reviewers for their helpful comments on earlier versions of this manuscript. We also thank Bryan Burt for running the control participants in Kingston, Canada. This research was funded by the Japan Society for the Promotion of Science and the Ministry of Education, Culture, Sports, Science and Technology.

[1]Attempts were made to eliminate allophonic differences such as vowel devoicing within each minimal pair by asking listeners to read the minimal pairs together in sets rather than separately.

[2]As mentioned previously, inclusion versus exclusion of EC01's data did not alter the main results.

Amano, S., and Kondo, T. (**2000**). *Nihongo-no Goi-Tokusei (Lexical Properties of Japanese)* (Sanseido, Tokyo).

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/. IV. Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**, 2299–2310.

Fujisaki, H., Nakamura, K., and Imoto, T. (**1975**). "Auditory perception of duration of speech and non-speech stimuli," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 197–219.

Greenhouse, S. W., and Geisser, S. (**1959**). "On methods in the analysis of profile data," Psychometrika **24**, 94–112.

Hirata, Y. (**2004a**). "Effects of speaking rate on the vowel length distinction in Japanese," J. Phonetics **32**, 565–589.

Hirata, Y. (**2004b**). "Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts," J. Acoust. Soc. Am. **116**, 2384–2394.

Hirata, Y., and Lambacher, S. G. (**2004**). "Role of word-external contexts in native speakers' identification of vowel length in Japanese," Phonetica **61**, 177–200.

Hirata, Y., Whitehurst, E., and Cullings, E. (**2007**). "Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates," J. Acoust. Soc. Am. **121**, 3837–3845.

Hirata, Y., and Whiton, J. (**2005**). "Effects of speaking rate on the single/geminate stop distinction in Japanese," J. Acoust. Soc. Am. **118**, 1647–1660.

Huggins, A. W. F. (**1972**). "Just noticeable differences for segment duration in natural speech," J. Acoust. Soc. Am. **51**, 1270–1278.

Ikuma, Y., and Akahane-Yamada, R. (**2004**). "An empirical study on the effects of acoustic and semantic contexts on perceptual learning of L2 phonemes," Annual Review of English Language Education in Japan **15**, 101–108.

Kaiki, N., and Sagisaka, Y. (**1992**). "The control of segmental duration in speech synthesis using statistical methods," in *Speech Perception, Production, and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, (Ohmsha, Tokyo), pp. 391–402.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**2002**). "Effects of phoneme class and duration on the acceptability of temporal modifications in speech," J. Acoust. Soc. Am. **111**, 387–400.

Kinoshita, K., Behne, D., and Arai, T. (**2002**). "Duration and F0 as perceptual cues to Japanese vowel quantity," in Proceedings of the 2002 International Conference on Spoken Language Processing, Denver, CO, pp. 757–760.

Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.

Kubozono, H. (**2002**). "Temporal neutralization in Japanese," in *Papers in Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner (Mouton, Berlin).

Lenneberg E. (**1967**). *Biological Foundations of Language* (Wiley, New York).

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). "Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new phonetic categories," J. Acoust. Soc. Am. **94**, 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., and Yamada, T. (**1994**). "Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories," J. Acoust. Soc. Am. **96**, 2076–2087.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**, 874–886.

Minagawa-Kawai, Y., Maekawa, K., and Kiritani, S. (**2002**). "Effects of pitch accent and syllable position in identifying Japanese long and short vowels: Comparison of English and Korean speakers," Journal Phonetic Soc. of Japan **6**, 88–97 (in Japanese).

Oguma, R. (**2000**). "Perception of Japanese long vowels and short vowels by English-speaking learners," Japanese-Language Education around the Globe **10**, 43–54 (in Japanese).

Omuro-Hayashida, K. (**1999**). "Pitch or duration? The perception of morae in long vowels in Japanese: A comparison between Japanese and English native speakers," in *Transactions of the Technical Committee on Psychological and Physiological Acoustics* [Acoustical Society of Japan (in Japanese), Kumamoto, Japan], Vol. **29**, pp. 1–8.

Sagisaka, Y., and Tohkura, Y. (**1984**). "Phoneme duration control for speech synthesis by rule," Trans. Inst. Electron., Inf. Commun. Eng. A **J67-A**, 629–636.

Tabuchi, S., Tokiyoshi, S., Yamakawa, K., Kai, T., Baba, R., Ueno, K., Usagawa, T., and Ebata, M. (**1997**). "Japanese long vowels: About the role of the pitch in morae perception," in *Transactions of the Technical Committee on Psychological and Physiological Acoustics* [Acoustical Society of Japan (in Japanese), Kumamoto, Japan], Vol. **27**, pp. 1–8.

Tajima, K., and Erickson, D. (**2001**). "Syllable structure and the perception of second language speech," in *Bunpo to Onsei 3 (Speech and Grammar 3)*, edited by Spoken Language Working Group (Kuroshio, Tokyo), pp. 221–239.

Tajima, K., Kato, H., Rothwell, A., and Munhall, K. G. (**2003a**). "Native and non-native perception of moraic phonemes in Japanese: Effect of identification training and exposure," in Proceedings of the 2003 Spring Meeting of the Acoustical Society of Japan, Tokyo, pp. 491–492.

Tajima, K., Kato, H., Rothwell, A., and Munhall, K. G. (**2003b**). "Perception of phonemic length contrasts in Japanese by native and non-native listeners," in Proceedings of the 15th International Congress of Phonetics Sciences, Barcelona, Spain, pp. 1585–1588.

Takeda, K., Sagisaka, Y., and Kuwabara, H. (**1989**). "On sentence-level factors governing segmental duration in Japanese," J. Acoust. Soc. Am. **86**, 2081–2087.

Toda, T. (**2003**). *Second Language Speech Perception and Production: Acquisition of Phonological Contrasts in Japanese* (University Press of America, Lanham, MD).

Uchida, T. (**1998**). "Categorical perception of relatively stead-state speech sound duration in Japanese moraic phoneme," Journal Phonetic Soc. of Japan **2**, 71–86 (in Japanese).

Wang, Y., Spence, M. V., Jongman, A., and Sereno, J. A. (**1999**). "Training American listeners to perceive Mandarin tones," J. Acoust. Soc. Am. **106**, 3649–3658.

Warner, N., and Arai, T. (**2001**). "The role of the mora in the timing of spontaneous Japanese speech," J. Acoust. Soc. Am. **109**, 1144–1156.

Watson, C., Kelly, W., and Wroton, H. (**1976**). "Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty," J. Acoust. Soc. Am. **60**, 1176–1186.

Yamada, R. A., (**1995**). "Age and acquisition of second language speech sounds: Perception of American English /r/ and /l/ by native speakers of Japanese," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD), pp. 305–320.

Yamada, T., Yamada, R. A., and Strange, W. (**1994**). "Perceptual learning of Japanese mora syllables by native speakers of American English: An analysis of acquisition processes of speech perception in second language learning," in Proceedings of the 1994 International Conference on Spoken Language Processing, (Yokohama, Japan), pp. 2007–2010.