Cross-language perceptual similarity predicts categorial discrimination of American vowels by naïve Japanese listeners

Winifred Strange^{a)} and Miwako Hisagi

*PhD Program in Speech, Language, Hearing Sciences, The City University of New York-Graduate School and University Center, 365 Fifth Avenue, New York, New York 10016
strangepin@aol.com; mhisagi@hotmail.com*

Reiko Akahane-Yamada

*ATR Intelligent Robotics and Communication Laboratories, ATR Learning Technology Corporation,
2-2-2 Hikaridai, Seika-cho, Soraku, Kyoto, Japan, 619-0288
yamada@atr-it.jp*

Rieko Kubo

*Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1291 Japan
rkubo@jaist.ac.jp*

Current speech perception models propose that relative perceptual difficulties with non-native segmental contrasts can be predicted from cross-language phonetic similarities. Japanese (J) listeners performed a categorial discrimination task in which nine contrasts (six adjacent height pairs, three front/back pairs) involving eight American (AE) vowels [i:, ɪ, ε, æ:, ɑ:, ʌ, ʊ, u:] in /hVbə/ disyllables were tested. The listeners also completed a perceptual assimilation task (categorization as J vowels with category goodness ratings). Perceptual assimilation patterns (quantified as categorization overlap scores) were highly predictive of discrimination accuracy ($r_s = 0.93$). Results suggested that J listeners used both spectral and temporal information in discriminating vowel contrasts.

© 2011 Acoustical Society of America

PACS numbers: 43.71.Hw, 43.71.Es, 43.71.An [JH]

Date Received: May 19, 2011 Date Accepted: July 28, 2011

1. Introduction

Current theories of cross-language and second-language (L2) speech learning postulate that perceptual differentiation of non-native segmental contrasts is strongly influenced by the phonetic similarity of non-native phones to native (L1) phonological categories (Best, 1995, Best and Tyler, 2007; Flege, 1995; Strange and Shafer, 2008). Empirically, similarity relations are established using a perceptual assimilation (PA) task in which listeners categorize non-native phones as exemplars of native categories and judge their “goodness” (Strange, 2007). In Best’s (1995) Perceptual Assimilation Model (PAM), L2 phones may be consistently categorized (C) or uncategorized (U), i.e., not consistently assimilated to a single L1 category. If two L2 phones are consistently assimilated as equally good exemplars of a single L1 category [single-category (SC) pattern] they will be very difficult to discriminate; if assimilated to two different L1 categories [two-category (TC) pattern], they will be discriminated easily. Intermediate levels of difficulty are predicted for L2 contrasts in which only one phone is consistently categorized [categorized-uncategorized (CU) pattern] or if both are assimilated to the same native category, but one is judged a better exemplar [category-goodness (CG) pattern].

^{a)} Author to whom correspondence should be addressed.

The automatic selective perception model (Strange and Shafer, 2008; Strange, 2010) posits that naive listeners perceive non-native segments/sequences on the basis of well-learned selective perception routines (SPRs). In processing continuous speech, L1 SPRs are attuned to the most reliable information for distinguishing phonologically contrastive segments in the L1, whereas phonetic details that are not critical for distinguishing L1 lexical items are ignored (phonological mode of processing). However, depending on stimulus complexity and task demands, listeners may be able to access phonetic details that are typically ignored. If stimuli are simple and the task allows for focal attention to be placed on intra-category acoustic/articulatory differences, listeners will reveal a phonetic mode of processing, in which deviations from native patterns may be detected. In a PA task with simple materials, naive listeners will be able to make fine-grained judgments about context-specific L1/L2 phonetic (dis)similarities (Strange *et al.*, 2009).

In the present study, naïve Japanese listeners (living in Japan) completed two tests: (1) a categorial AXB discrimination test in which American English (AE) vowels [i:, ɪ, ɛ, æ:, ɑ:, ʌ, ʊ, u:] were arranged into 13 contrasting pairs (see Sec. 2 and Table 2) and (2) a PA task in which the AE stimuli were categorized as exemplars of the nearest Japanese categories and rated for “goodness” on a nine-point Likert scale. Discrimination performance across the 13 pairs was then compared with perceptual assimilation results to see if assimilation patterns predicted relative accuracy. Of particular interest was the extent to which duration differences of intrinsically short, mid, and long AE vowels influenced PA patterns and discrimination accuracy.

Japanese (J) contrasts five spectrally distinct vowels [i, e, a, o, u] that are also distinguished phonologically by length (one-mora vs two-mora, indicated here by dual symbols, [ii...uuu]) and palatalization of [ʲa(a), ʲo(o), ʲu(u)]. Duration ratios of two- vs one-mora vowels are at least 2 to 1 in citation-form materials and greater than 1.5 to 1 in continuous speech (Nishi *et al.*, 2003). AE contrasts 11 (non-rhotic) vowels: long [eː, æː, ɑː, ɔː, ɔː, oː], intermediate [ɪː uː], and short [ɪ, ɛ, ʌ, ʊ]; long/short duration ratios are approximately 1.2–1.5 to 1 in citation-form materials and vary across speakers and phonetic contexts (Strange *et al.*, 2002).

Past research on perceptual assimilation of AE vowels by Japanese listeners (Strange *et al.*, 1998, 2001) showed that short AE vowels in citation-form disyllables were consistently assimilated as one-mora J vowels (94%), but intermediate and long vowels were not consistently assimilated as two-mora categories (42%). Spectral

Table 1. Perceptual assimilation test results, summed over all participants.

AE Vowel	Modal J Category	Percent modal Responses ^a	Consistent Subjects ^b	Median Rating ^c	Other response Categories ^d (%)	Consistent Subjects ^e
i:	hii	73	9	7	hi (25)	1
ɪ	he	97	21	7		
ɛ	h ^ʲ a	73	10	3	he (16); ha (9)	ha 1
æ:	haa	58	2	3.5	ha (21) h ^ʲ aa (11); h ^ʲ a (8)	
ɑ:	haa	96	19	6		
ʌ	ha	91	18	6		
ʊ	huu	57	0	4	huuu (30); ho (7)	
u:	huuu	50	5	6	huu (49)	5

^aPercentage of total trials (summed over 21 participants) the most frequent response was selected.

^bNumber of participants who selected the J modal response on at least 13 of 15 trials.

^cMedian rating on modal response trials, summing over all participants.

^dOther J responses receiving at least 3% of overall responses (summed over participants).

^eNumber of participants who selected a minority response on at least 13 of 15 trials.

Table 2. Perceptual assimilation patterns, perceived similarity overlap, and discrimination accuracy for 13 contrasts among 8 AE vowels.

Contrast type	Contrast	PAM patterns ^a	Overlap	Percent	Median (range)	
	Pair	(Feature) ^b	Score (%)	Correct ^c	Correct ^d	
Height + length	i:/ɪ	TC (H); CG (L)	0	99	12	(11–12)
	ɛ/æ:	UC (H); CU (L)	19	82	10	(8–12)
	ʌ/ɑ:	SC (H); TC (L)	7	90	11	(6–12)
Height only	u:/ʊ ^e	SC (H); UU (L)	79	71	9	(5–11)
	ɪ/ɛ	CU (H); SC (L)	16	88	11	(8–12)
	ʊ/ʌ	TC (H); UC (L)	10	98	12	(10–12)
Front/back	æ:/ɑ:	CG (F/B); UC(L)	61	76	9	(4–12)
	ɛ/ʌ	UC (F/B); SC (L)	16	94	12	(9–12)
	ɪ/ʊ	TC (F/B); CU (H)	0	99	12	(10–12)
Control pairs		TC (H)	0	98–100	12	(10–12)

^aPAM patterns: TC = two-category; CG = category-goodness; UC = uncategorized-categorized; CU = categorized-uncategorized; SC = single-category.

^bH = height (F1); L = length (duration); F/B = front-back position (F2).

^cMean correct responses on the AXB discrimination task expressed, as a percentage of total opportunities (12 trials/contrast × 21 listeners).

^dMedian and range of scores on each pair (out of 12 possible).

^eAlthough transcribed as long/short vowels, mean durations for these tokens did not differ.

assimilation patterns showed that, whereas AE vowels [i: ɑ:, u:, ʊ] were categorized very consistently as one- or two-mora J [i, ɑ, u, ʊ], respectively, AE [ɪ, ɛ, æ:, ʌ,] were less consistently categorized. No discrimination tests of any AE vowel pairs were administered. Thus, the present study extends previous work to a new speaker and dialect of AE, tests categorical discrimination of many contrasts, and correlates relative discrimination difficulties with PA patterns.

2. Method

Participants were 21 young adult (10 males) native speakers of Kansai dialect (age 19–25 yr) with little English conversational exposure and standard English instruction in school; all passed a hearing screening (500, 1K, 2K, 4K Hz at 20 dB hearing level). They were tested at the Department of Multilingual Learning, ATR Human Information Science Laboratories and paid for their participation.

Three tokens of each of eight [hVbə] disyllables, produced in citation form by a young adult male speaker of New York dialect served as stimuli. Figure 1(A) shows formant values measured at the 50% point of each vocalic nucleus. Figure 1(B) shows the formant trajectories over the middle half of each syllable nucleus (arrowheads indicate 75% points). Figure 1(C) shows the mean vocalic durations. Note that [u/ʊ] did not differ in duration as expected.

All Japanese participants completed an AXB Categorical Discrimination test, followed by a Perceptual Assimilation (PA) test on the same day. Stimuli were presented via PC computer binaurally over (Stax Professional) headphones at a preset comfortable listening level, using customized software for stimulus delivery and response registration. For the AXB task, 13 pairs were contrasted: four control pairs [i:/ɛ, u:/ʌ, æ:/ɪ, ɑ:/ʊ] and nine experimental pairs—six adjacent height pairs [i:/ɪ, u/ʊ, ɪ/ɛ, ʊ/ʌ, ɛ/ɑ:, ʌ/ɑ:] and three front-back pairs [ɪ/ʊ, ɛ/ʌ, æ:/ɑ:]; three adjacent height pairs

Fig. 1. Acoustic parameters of 24 stimuli (3 tokens of each vowel): (A) mid-syllable first (F1) and second (F2) formant values (in Bark); (B) mean formant trajectories (25–50–75%) in F1/F2 Bark space; (C) mean durations.

also contrasted in intrinsic duration. Stimuli were combined such that all three members of each AXB triad were acoustically different tokens. The three tokens of each type were arranged in four combinations of AXB triads (500 ms ISI) for a total of 12 trials for each contrast. There were seven blocks of 26 trials (two trials $\times$ 13 contrasts); the first block was used for stimulus/task familiarization and was not scored. The total length of the test was approximately 20 min.

For the PA test, the stimuli were arranged into 6 randomized blocks of the 24 stimuli (eight vowels, three tokens each). The first block was used for task familiarization and calibration of rating judgments and was not scored. A stimulus was presented, then 16 [hV(V)] response alternatives appeared (Kana characters); after each categorization response, the same stimulus was presented again, and a nine-point Likert scale appeared (1 = foreign-like; 9 = Japanese-like) on which the listener judged the token's goodness. A total of 15 judgments (three tokens, five repetitions each) for each of the eight AE vowels was scored. The length of this test was 20–30 min.

3. Results

Although completed second, the PA results are presented first, followed by the AXB discrimination results. PA responses for each AE vowel were tallied within and across listeners and the group modal response, followed by other responses of greater than 5% of total responses were determined. In addition, the number of J listeners who consistently categorized each AE vowel (at least 13 out of 15 opportunities) as a particular J category was tallied (see Table 1).

Overall, the most consistently assimilated AE vowels with respect to both spectral and temporal features were [ɪ, ʌ, ʌ]. AE [i:, æ:, ʊ, u:] were each consistently assimilated to only one of the five spectral categories, but categorization as one- or two-mora J vowels was inconsistent both within and across J listeners, and [æ:] was categorized as more similar to a palatalized vowel on a substantial number of trials. Finally, [ɛ] was consistently assimilated as [h^ha] by about half the listeners, but was uncategorizable for others (responses spread among [h^ha, he, ha]).

Using the group patterns of perceptual assimilation, the 13 contrasting pairs were categorized in terms of PAM patterns, and in terms of overlap scores (Levy, 2009). For the latter, percentages of minority responses that overlapped across each contrasting pair were summed. For [æ:/ɑ:] where the group modal response was the same, the 58% [haa] responses for [æ:] were added to the 3% [ha] responses for [ɑ:] = 61%. PAM patterns and overlap scores for the nine experimental contrasts are presented in Table 2, along with the group results on the AXB discrimination test.

As expected, all four control contrasts yielded TC assimilation patterns, with no overlap in category responses. Discrimination of these contrasts was at ceiling (median = 12/12 correct for all four pairs). Four of the experimental contrasts [i:/ɪ, ʊ/ʌ, ɛ/ʌ, ɪ/ʊ] also showed ceiling performance, predictable from PAM patterns and small overlap scores. Another three pairs [ɪ/ɛ, ɑ:/ʌ, ɛ/æ:] were discriminated somewhat less accurately overall, with large inter-listener variability in performance. Group PAM patterns were CU/UC/SC on height or front/back features, with overlap scores of 7–19%. The two most difficult contrasts [æ:/ɑ:, ʊ/u:] were discriminated just above chance levels on average (median = 9) with seven and eight (of 21) participants, respectively, performing no better than chance. These pairs showed group SC patterns or small CG differences, and considerably larger overlap scores.

Because it was difficult to rank order pairs by PAM patterns on dual features, the relationship between perceived similarity and discrimination performance was assessed using group overlap scores and group percent correct scores (see Table 2). A Spearman rank-order correlation of the nine experimental pairs yielded a very high coefficient ($r_s = 0.937$), supporting the claim that PA patterns are predictive of relative discrimination difficulties of non-native vowel contrasts by naïve listeners (> 80% of variance accounted for).

4. Discussion and conclusions

The results of this study demonstrated that some AE vowels are very difficult for naïve Japanese listeners to perceptually differentiate, even when the stimuli are simple and the task relatively undemanding (short ISI, AXB format) which allows listeners to focus attention on detailed phonetic information that distinguishes non-native vowel categories. In general, the results support the prediction that vowel duration ratios of well less than 2-to-1 served to differentiate some AE contrasts for J listeners. When duration differences were not available and spectral differences were relatively small, performance suffered, especially for [æ:/ɑ:], and, for this speaker, [ʊ/u] [see Fig. 1(C)]. The other three pairs that were not temporally differentiated [ɪ/ɛ, ʊ/ʌ, ɛ/ʌ] were discriminated well above chance by most listeners on the basis of perceived height, front-backness, and/or perceived palatalization.

The overall pattern of perceptual assimilation shown here differs from those reported for productions of AE speakers of Midwestern dialect (Strange *et al.*, 1998)

for four vowels: (1) AE [ɪ] was consistently assimilated to J [e] (vs more often to J [i] in the prior study), (2) AE [ɛ] was assimilated more often to J [ʔa, a] (vs more often to [e]), (3) AE [æ:] was more consistently assimilated to J [a(a), ʔa(a)] (vs more often to [e(e)]), (4) AE [ʌ] was more consistently assimilated to J [a] (responses spread over J [a, o, u] in the previous study). See also Japanese listeners' assimilation of Australian English vowels (Bundgaard-Nielsen *et al.*, 2011).

The extent to which these differences are due to dialect variation, individual speakers, or listener differences cannot be determined without further comparisons with multiple speakers of both dialects using the same listeners. However, it suggests that the perceptual assimilation of these non-native AE vowels by Japanese learners may be dependent on their specific phonetic realization. Therefore, before making generalizations about the relative difficulty of AE contrasts involving these vowels, it will be necessary to establish cross-language perceptual similarity patterns for specific listeners and speakers. Furthermore, since perceptual assimilation patterns differ as a function of whether listeners are presented citation or sentence materials (Strange *et al.*, 1998, 2001), the present results provide only a baseline for relative discrimination accuracy when listeners can invoke a phonetic mode of processing (Strange, 2010).

Acknowledgments

Supported by NIDCD Grant No. 00323 to W.S. and by a grant from ATR to R.A.-Y.

References and links

- Best, C. T. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York, Timonium, MD), pp. 171–204.
- Best, C. T., and Tyler, M. D. (2007). "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, edited by O.-S. Bohn and M. J. Munro (John Benjamins, Amsterdam), pp. 13–34.
- Bundgaard-Nielsen, R. L., Best, C. T., and Tyler, M. D. (2011). "Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories," *Appl. Psychol.* **32**, 51–67.
- Flege, J. E. (1995). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York, Timonium, MD), pp. 233–277.
- Levy, E. S. (2009). "On the assimilation-discrimination relationship in American English adults' French vowel learning," *J. Acoust. Soc. Am.* **126**, 2670–2682.
- Nishi, K., Akahane-Yamada, R., Kubo, R., and Strange, W. (2003). "Acoustic comparisons of Japanese and English vowels produced by native speakers of Japanese," *J. Acoust. Soc. Am.* **114**, 2364.
- Strange, W. (2007). "Cross-language phonetic similarity of vowels: Theoretical and methodological issues," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, edited by O.-S. Bohn and J. M. Munro (John Benjamins, Amsterdam), pp. 35–55.
- Strange, W., Weber, A., Levy, E., Shafiro, V., and Nishi, K. (2002). "Within- and across-language acoustic variability of vowels spoken in different phonetic and prosodic contexts: American English, North German, and Parisian French," *J. Acoust. Soc. Am.* **112**, 2384.
- Strange, W. (2010). "Automatic selective perception (ASP) of first and second language speech: A working model," *J. Phonetics*, doi: 10.1016/j.wocn.2010.09.001.
- Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., and Nishi, K. (2001). "Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners," *J. Acoust. Soc. Am.* **109**, 1691–1704.
- Strange, W., Levy, E. S., and Law II, F. F. (2009). "Cross-language categorization of French and German vowels by naive American listeners," *J. Acoust. Soc. Am.* **126**, 1461–1476.
- Strange, W., and Shafer, V. L. (2008). "Speech perception in second language learners: The re-education of selective perception," in *Phonology and Second Language Acquisition*, edited by J. G. Hansen Edwards and M. L. Zampini (John Benjamins, Amsterdam), pp. 153–191.
- Strange, W., Yamada, R. A., Kubo, R., Trent, S. A., Nishi, K., and Jenkins, J. J. (1998). "Perceptual assimilation of American English vowels by Japanese listeners," *J. Phonetics* **26**, 311–344.