PERCEPTUAL ASSIMILATION OF AMERICAN ENGLISH VOWELS BY JAPANESE LISTENERS

Winifred Strange[1], Reiko Akahane-Yamada[2], Brett. H. Fitzgerald[1], Rieko Kubo[3]

1. University of South Florida, 4202 E. Fowler Avenue, Tampa, FL, 33620-8150, USA
2. ATR Human Information Processing Research Laboratories, Seika-cho, Soraku-gun, Kyoto 619-02, Japan; 3. Nara, 631, Japan

ABSTRACT

To assess cross-language patterns of perceptual assimilation directly, 24 Japanese (J) listeners were presented American English (AE) vowels produced in citation-form /hVbɑ/ bisyllables and in a carrier sentence by 4 adult male speakers. They selected the J *katakana* character(s) which contained the vowel most similar to the AE vowel and rated the goodness-of-fit on a 7-point scale. Results showed that, as expected, AE vowels were judged as most similar to J vowels which were adjacent in acoustic-articulatory "vowel space." However, the consistency of assimilation and goodness of fit of AE vowels to J categories varied considerably with speech style (citation vs sentence). Assimilation of long AE vowels to two-mora response categories was much more consistent for target syllables produced in sentences. Acoustical analysis of stimuli suggested that listeners judged the duration of target vowels in citation bisyllables with respect to the following (constant) vowel. Other differences in perceptual assimilation patterns as a function of speech style could not easily be attributed to differences in speakers' productions. These results have implications for theories of L2 speech learning and for training studies of non-native speech sounds.

1. INTRODUCTION

Patterns of perceptual assimilation of non-native phones to native phonetic categories have been hypothesized to be predictive of difficulties adult second-language (L2) learners have in learning to perceive and produce non-native phonetic categories (Best, 1995; Flege, 1992). However, little research has *directly* assessed cross-language perceptual assimilation patterns by second-language learners. This study is part of a larger project which directly assesses the perceptual assimilation of non-native vowels to native phonetic categories in cross-language comparisons of Japanese, American English, and German. Of particular interest is the extent to which perceptual assimilation patterns are influenced by contextual variables such as speech style (citation vs sentence) and consonantal context. A previous study of American English (AE) listeners' perceptual assimilation of North German (NG) vowels (Trent, et al., 1995) found that patterns of assimilation varied with speech style. In particular, listeners utilized relative vowel duration information more effectively when vowels were presented in sentence context.

In the present study, patterns of perceptual assimilation of American English (AE) vowels by native speakers of Japanese (J) were investigated. The J vowel inventory is described as consisting of five monophthongal vowels /i, e, a, o, u/ which differ in tongue height and position (vowel quality). Vowel length is also contrastive in J; thus, long (two-mora) and short (one-mora) versions of each vowel are distinctive. In comparison, the AE vowel inventory consists of 11 vowels which vary in vowel quality, while vowel length is considered redundant (although vowels do vary systematically in intrinsic duration). In this study, the AE vowels were produced in /hVbɑ/ syllables spoken as citation form utterances (in lists) and in a carrier sentence, "I say the /hVb/ on the tape." J listeners were asked to categorize each of 11 AE vowels in the /hV/ syllables as "most similar" to one of 18 Japanese (J) one-mora or two-mora response alternatives (5 short vowels, 5 long vowels and 8 two-vowel combinations) and to rate the "goodness-of-fit" of the AE stimulus to that J category. Three questions were asked:

(1) To what extent are perceptual assimilation patterns predictable on the basis of cross-language phonetic similarities in (a) vowel quality (proximity in articulatory/acoustic vowel space) and (b) vowel length/ duration?

(2) Do assimilation patterns vary with speech style (citation vs sentence) ? Specifically, (a) do perceptual categories "shift" in vowel space across context conditions, and (b) are intrinsic vowel duration differences utilized more effectively in categorizing target syllables presented in sentence context?

(3) Are the differences in perceptual assimilation patterns as a function of speech style accounted for by (a) differences in speakers' productions (F1/F2 shifts and/or relative durational differences) or by (b) differences in perception (listeners' assimilation strategies). or both?

2. METHOD

2.1 Subjects

Twenty-four young adult native speakers of Japanese from the Kansai area served as subjects. The 13 females and 11 males ranged in age from 18 to 23 years old. They had all received "standard" instruction in English, which consists of 6 years of schooling. Reading and writing skills are emphasized with little or no conversational experience with native speakers of English. None of the subjects had spent an extended period of time in an English speaking country.

2.2 Stimulus Materials

Four young adult male native speakers of American English (AE) produced the stimulus corpus. None displayed a strong regional dialect, as determined by the first author; all speakers differentiated the AE vowels [ɑ:] and [ɔ:] Three speakers resided in Florida at the time of recording; the fourth speaker resided in Japan, but spoke English almost all of the time. Each speaker produced 4 tokens of each of 11 vowels [i:, ɪ, e͜ɪ, ɛ, æ:, ɑ:, ʌ, ɔ:, o͜ʊ, ʊ, u:] in each context. The order of vowels was randomized across repetitions. The first repetition in each context was not used unless the remaining three contained extraneous noise. dysfluencies, or an inappropriate pitch contour, as judged by the first author. In those few cases, the first repetition of the vowel was utilized. Thus, a total of 264 stimuli were used: 11 vowels x 3 tokens x 2 contexts x 4 speakers. Stimuli were recorded using a digital audio tape

(DAT) recorder at a 44.1 kHz sampling rate and transferred to computer files with downsampling to 22.05 kHz.

2.3 Procedures

Subjects completed cross-language phonetic categorization and goodness ratings of the stimuli using the following procedure. After presentation of each stimulus (bisyllable or sentence), subjects categorized the /hV/ target syllable as "most similar" to 1 of 18 J /hV(V)/ response alternatives, displayed in *katakana* characters. After the second presentation of the same stimulus, the subject rated its "goodness-of-fit" to the chosen alternative on the scale from 1 to 7 (7 = best fit). All subjects were tested individually (in Japan) using an interactive computer program. Stimuli were presented via earphones at a comfortable listening level. Subjects could repeat a stimulus or change their categorization response, but were discouraged from doing so.

A repeated-measures design was used in which each listener was presented the stimuli of all 4 speakers' productions in both speech styles. Citation and sentence utterances were presented in separate sessions on different days, with order counterbalanced across subjects. Subjects completed all four speaker conditions in each session, with order of speakers counterbalanced across listeners in a pseudo-Latin square design. For each speaker, subjects completed a 33-item familiarization block, then completed the 99-item test (3 randomized blocks of the 33 utterances) for a total of 9 judgments on each vowel by each speaker.

Data are reported in terms of the percentage of times each AE vowel was categorized as most similar to a J response category, summed over all listeners. In addition, the median rating assigned to the responses was calculated.

3. PERCEPTUAL RESULTS

AE	J	Resp. 1 (%)	Median Rating	J	Resp. 2 (%)	Median Rating
i:	i	59	6	ii	40	6
eɪ	eɪ	65	5	e	16	4
æ:	a	30	2	e	29	2
a:	a	79	6	aa	20	5
ɔ:	o	31	3	oo	28	4
oʊ	o	54	5	ou	27	5
u:	u	61	5	uu	39	5
ɪ	i	58	3	e	39	4
ɛ	e	83	3	a	9	2
ʌ	a	64	4	u	18	1
ʊ	u	83	3	uu	14	3

Table 1: Modal (Resp 1) and next most frequent response (Resp 2) alternatives (percentages of opportunities) and median goodness ratings for 11 AE vowels: Citation-form bisyllables.

Tables 1 and 2 present the overall results for citation and sentence conditions, respectively, summing over listeners and speakers within each condition The AE and J response alternatives are given in IPA symbols; for J categories a single symbol represents the *katakana* syllable containing a short vowel (one-mora syllable), a double symbol (e.g. ee) represents the *katakana* syllable containing a long vowel (two-mora syllable), and a vowel combination represents the two-symbol (two-mora) sequence in which the vowel differs (hV1 + V2). In these tables, only the most frequently selected responses (columns 2-4) and the second most frequently selected responses (columns 5-7) are presented. (For all AE vowels, responses were distributed across more than two alternatives in one or both contexts; thus, percentages do not add to 100%.) The intrinsically long and diphthongized AE vowels are given in the first 7 rows, followed by the 4 intrinsically short vowels.

AE	J	Resp. 1 (%)	Median Rating	J	Resp. 2 (%)	Median Rating
i:	ii	83	6	i	17	6
eɪ	ei	78	5	ii	15	4
æ:	aa	34	2	ee	16	2
a:	aa	71	5	a	21	3
ɔ:	oo	50	4	aa	26	2
oʊ	ou	54	4.5	oo	18	5
u:	uu	87	5	u	13	4
ɪ	i	77	4	e	16	4
ɛ	e	58	3	a	25	2
ʌ	a	65	4	u	15	3
ʊ	u	53	3.5	uu	42	3

Table 2: Modal (Resp 1) and next most frequent response (Resp 2) alternatives (percentages of opportunities) and median goodness ratings for 11 AE vowels: Sentences.

In general, a comparison of AE vowels (column 1) with modal response categories (column 2) suggests that AE vowels in both citation and sentence utterances were assimilated most often to J vowel categories which were most similar in articulatory/acoustic "vowel space." AE low front and back [æ:] and [ɑ:] assimilating to J low [a] or [aa] and AE mid-low and mid back [ɔ:] and [oʊ] assimilating to J [o] , [oo] or [ou]. The low-mid central [ʌ] assimilated primarily to J [a] in both citation and sentence conditions. However, the overall consistency in assimilation varied considerably across vowels, with percentages of modal responses ranging from 30% to 87% (mean = 61% and 65% for citation and sentence conditions, respectively). Goodness-of-fit ratings also indicated that some AE vowels were perceived as much more similar to J vowels than others, with median ratings ranging from 6 to 2 in both context conditions. When duration is not considered, (i.e., summing across one-mora and two-mora response categories) the vowels [i:, ɑ:, and ʊ, and u:] were categorized with greater than 90% consistency in both contexts. In contrast, the vowels [æ:, ɔ:, ʌ] were rather poorly assimilated to any one J vowel "quality" in both contexts.

Turning to the question of the effects of speech style (citation vs sentence) on assimilation patterns, it is immediately apparent that differences in intrinsic duration of AE vowels were utilized more consistently in perceptual assimilation of vowels produced in sentence context. Long AE vowels in sentence utterances were assimilated to J two-mora response in citation utterances. Indeed, for the latter, the modal response was a one-mora category for 6 of the 7 vowels. Alternatively, short AE vowels tended to be assimilated to one-mora response alternatives somewhat less well in the sentence condition than in the citation condition, suggesting a response bias for short responses in the citation condition (see below). Nevertheless, if "accuracy" is defined as assimilation of AE long and diphthongized vowels to J two-mora response categories and AE short vowels to one-mora categories, then utilization of intrinsic duration information was considerably more accurate for sentence utterances (84% correct overall) than for citation utterances (61% correct overall).

For 4 AE vowels, speech style also influenced the assimilation patterns with respect to "vowel quality." AE [ɪ] was assimilated to J [i] more consistently in the sentence condition while AE [ɛ] was assimilated to J [e] more consistently in the citation condition. Assimilation of AE [ɔ:] and [eɪ] also varied across conditions, especially in the second most frequent responses.

4. Acoustical Analysis

In order to answer the third question posed in the introduction, acoustical analysis of the stimulus corpora was undertaken. Formant frequencies at target syllable midpoint, formant trajectories over the middle half of each target syllable, and target syllable vocalic durations were measured. Of interest was the extent to which these acoustical parameters varied with context. Since the most striking overall perceptual differences were in the assimilation on the basis of intrinsic duration, absolute and relative durations of the vocalic portions of target syllables were compared, as shown in Table 3. Average durations of long and short vowels for each speaker's productions in each context are presented in columns 2 and 3. In addition, the relative durations of long vs short vowels are expressed as a ratio (long/short) in the third column. As this table indicates, relative durations were equivalent, on average, across speech styles. Thus, the poorer differentiation of citation utterances in terms of assimilation to one- and two-mora response categories cannot be explained by differences in the relative durations of the target syllables in the two conditions.

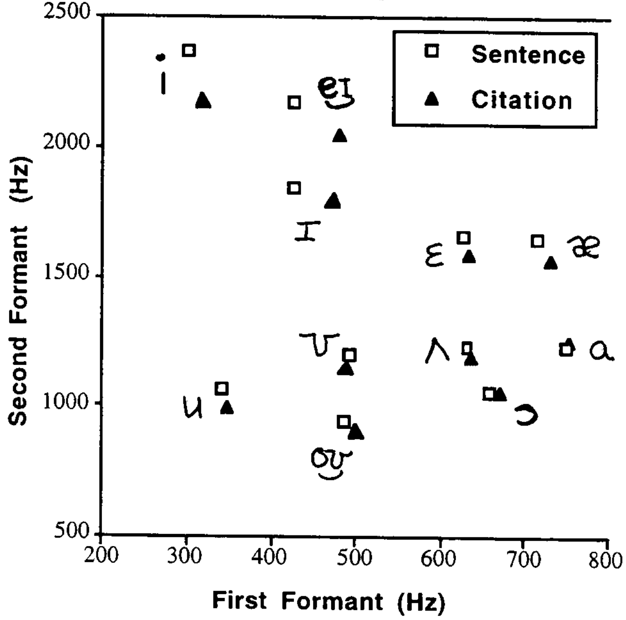
Speaker	Vowels Long	 Short	Ratio L/S
	Citation		
BF	132	94	1.41
KS	100	85	1.17
JM	126	99	1.27
MJ	101	81	1.25
Average	115	90	1.28
	Sentences		
BF	143	103	1.40
KS	126	103	1.23
JM	126	101	1.25
MJ	123	91	1.35
Average	129	99	1.31

Table 3: Durations (ms) of vocalic portions of target syllables in citation (above) and sentence (below) contexts.

When the durations of the target syllables were compared with the following [ɑ:] in citation utterances vs the [ɑ:] of "on" in sentence utterances, a striking difference emerged. For citation utterances, the average ratio of target syllables containing long vowels to the following syllable was 0.60, while the ratio of short vowel target syllables to [ɑ:] was 0.48. That is, even "long" vowels were short relative to the other vowel in the citation-form bisyllables. In contrast, duration ratios for short and long target syllables, relative to the following vowel in sentence utterances were 1.01 and 1.39 for short and long vowels, respectively. These differences in the immediate context, as well as the presence of the larger rhythmic context provided by the sentences, could account for the differences in assimilation patterns across contexts with respect to the utilization of intrinsic duration information. A follow-up study, in which the final vowel of the /hVbɑ/ syllables is shortened, is underway to examine these hypotheses.

Figure 1 presents the F1/F2 values (taken at syllable midpoint) for the target vowels produced in citation-form and sentence utterances, averaged over speakers and repetitions. In general, differences as a function of speech style in perceptual assimilation to different J vowel qualities for [ɪ, ɛ, ɔ:, eɪ] cannot be explained by differences in F1/F2 values at syllable midpoint for vowels produced in citation-form vs in sentences. While the more consistent assimilation of AE [ɪ] to J [i] might be explained by the lower average F1 values for this vowel in sentence context, perceptual shifts for the other vowels were not predictable from F1/F2 differences across contexts.

Figure 1: F1/F2 Midpoint Values



Formant trajectories across the middle half of the target syllables were compared to see if there were systematic differences in diphthongization of AE vowels as a function of speech style. Although patterns varied across speakers, systematic differences across contexts were not immediately apparent for the [ɪ, ɛ, ɔ:] which might account for differences in perceptual assimilation to J vowel qualities for these vowels.

For two of the speakers, formant trajectories for [eɪ] did vary in direction and extent across contexts. However, for one speaker movement toward [ɪ] was greater for sentence context (conforming to observed perceptual shifts) while for the other speaker, formant movement toward [ɪ] was greater for citation-form utterances. Detailed token-by-token analyses are necessary to further explore the extent to which these acoustic/articulatory differences correlated with differences in perceptual assimilation patterns

In summary, while relative durations of vowels in citation-form and sentence utterances did not vary, differences in the immediate and more distant *temporal context* of citation vs sentence utterances could account for differences in perceptual assimilation of long AE vowels to two-mora categories (and to a lesser extent short vowels to one-mora categories) across context conditions. Shifts in perceived vowel quality as a function of speech style, found for a few AE vowels, are not as clearly attributable to production differences. Further acoustical analyses and detailed examination of the correlation between perceptual and acoustic patterns of variation with speech style must be accomplished before a final answer to the third question can be attained.

5. Discussion

We can conclude from this study that Japanese listeners perceptually assimilate AE vowels to native phonetic categories using both "spatial" (vowel quality) information and temporal (relative duration) information. In general, AE vowels are assimilated to adjacent J categories in acoustic/articulatory "vowel space;" however, some AE vowels are assimilated to more than one J spatial category. AE short vowels are consistently assimilated to J short (one-mora) vowel categories. AE long vowels are not always categorized as most similar to long (2-mora) J vowels or two-vowel combinations. Speech style (citation vs sentence) significantly affected perceptual assimilation patterns, especially with respect to the utilization of relative duration differences. Acoustical analysis revealed that this may have been due to the temporal context in which the target syllables were embedded. Other differences in perceptual assimilation patterns across contexts could not be explained easily by differences in mid-syllable formant values or formant trajectories of target syllables

These results suggest that cross-language perceptual assimilation patterns are not context-independent. Theories which invoke cross-language phonetic similarity as predictors of L2 learning difficulty must take this into account. Contextual effects such as the ones shown here may account for why contrastive analyses of phoneme inventories are ineffective in predicting L2 learning difficulties. These data also suggest that the representations of L1 vowel categories may not be based on canonical (context-free) spectral and temporal values. Thus, simple comparisons across languages of the distribution of (isolated) vowels in an F1/F2 vowel space will not be adequate for predicting perceived phonetic similarity.

The findings reported here also have practical implications for foreign language teachers and researchers interested in improving the perception of non-native phonetic segments and contrasts. For instance, training studies aimed at improving Japanese listeners' perceptual differentiation of AE vowels should employ sentence-length materials to maximize subjects' utilization of vowel duration information.

REFERENCES

Best, C.T. (1995). A direct realist view of cross-language speech perception. In W. Strange (ed.) *Speech perception and linguistic experience: Issues in cross-language research.* Timonium, MD: York Press.

Flege, J. E. (1992). Speech learning in a second language. In C. Ferguson, et al. (eds.) *Phonological development: Models, research and application.* Timonium, MD: York Press.

Trent, S.A., Fitzgerald, B.H., Crouse, S.M., & Strange, W. (1995) Perceptual assimilation of North German vowels to American English categories. *Journal of the Acoustical Society of America, 97,* 3419 (abstract).

[Supported by NIH-NIDCD and ATR HIP Laboratories]