ACOUSTIC CORRELATES TO THE EFFECTS OF TALKER VARIABILITY ON THE PERCEPTION OF ENGLISH /r/ AND /l/ BY JAPANESE LISTENERS

James S. Magnuson[1] *and Reiko Akahane-Yamada*[2]

[1]Brain and Cognitive Sciences, University of Rochester [magnuson@bcs.rochester.edu]
[2]ATR Human Information Processing Research Laboratories, Kyoto [yamada@hip.atr.co.jp]

ABSTRACT

It is often reported that for non-native listeners of a language, some native speakers' productions of non-native contrasts are easier to understand than others' (e.g., [1]). However, these effects are not well-understood, as acoustic correlates to the effects have proven difficult to establish. We report analyses of subject differences and acoustic measurements which may help to describe the acoustic phenomena underlying one class of talker effects that we have reported previously; specifically, the interaction of *talker* and *talker condition* (the number of talkers heard within a block of trials -- one or several) [2]. Correlations between response measures and acoustic measures suggest that when stimuli from several talkers are mixed randomly in a block of trials, subjects without well-formed categories for /r/ and /l/ attempt to use the duration of the initial steady-state portion of an /r/ or /l/ stimulus (an unreliable cue) for categorization, whereas native speakers use F3 [6]. It also appears that they use this cue to establish criteria for "R"-"L" decisions, which they apply to the overall range of durations across all talkers in one block of trials.

1. INTRODUCTION

Differences in talker characteristics are a well-known source of problematic variability in speech perception. Due to differences in age, sex, size, dialect, and other factors, the way different talkers acoustically realize the same linguistic segments may be quite different, and the way they realize different linguistic segments may be quite similar [3]. All the same, while listening to their native languages, people have little trouble perceiving speech despite talker differences. Here, we examine the question of what effect talker differences have on the perception of non-native phonemes which contrast on dimensions that are not distinctive in the native language. The particular non-native contrast we discuss is American English (AE) /r/ and /l/ for adult, native speakers of Japanese.

Lively et al. [1] reported talker-specific differences in Japanese listeners' accuracy in /r/-/l/ identification training. The talker-specific patterns persisted even after extended training [4]. This suggests that talkers may give differential emphasis to the multiple cues to /r/ and /l/ (although acoustic correlates to the perceptual differences have not been previously reported), and that Japanese listeners have not experienced a sufficient sampling of the range of cues to /r/ and /l/ that occur across different talkers to be able to normalize for this kind of variability. Considering talker differences in cues as adding to the *range* of cues suggests a connection to previous work by Yamada and Tohkura [5, 6].

While the steady-state onset and frequency transition of F3 is sufficient for native speakers of AE to distinguish /r/ and /l/ across talkers, /r/ and /l/ also differ systematically in the spectral and temporal characteristics of the first two formants [7]. When Yamada and Tohkura [5] covaried spectral cues to /r/ and /l/ (F3 and F2), native speakers of AE clearly based their decisions on F3. In contrast, Japanese subjects' responses were influenced by both cues.

Yamada and Tohkura [6] also found that the range of variation in acoustic cues differentiating /r/ and /l/ presented in a block of trials affected the labeling performance of Japanese listeners, but not that of native speakers of AE. No matter what portion of a synthesized /r/ - to - /l/ continuum they presented to Japanese listeners, rates of "R" responses were approximately 50%. Given only the most /r/-like half of the continuum, only the most /l/-like half of the continuum, or the entire range, subjects apparently set criteria such that they responded "R" on approximately half the trials within a block. Yamada and Tohkura concluded that Japanese subjects without well-defined categories for /r/ and /l/ set criteria relative to the *range* of cues to /r/ and /l/ they heard within a block of trials. This contrasts with the nearly categorical criteria native speakers employed with the same stimuli (see [6] for details).

These results suggest that while Japanese subjects with limited experience in an English-speaking environment are able to divide a set of stimuli into "R" and "L" categories, they do not possess robust categorization criteria that they can apply to individual stimuli independently of cue variability. That is, they respond on the basis of *relative* differences in an ad-hoc fashion, rather than categorical criteria.

Considering the talker-specific accuracy patterns reported by Lively et al. [1] together with the range effects found by Yamada and Tohkura [6] we tested predictions about the effects of within-session talker variability [2]. If Japanese subjects attend mainly to inadequate cues (i.e., non-ecologically valid cues), one might predict that to naive Japanese listeners, the relative range of cues to /r/ and /l/ would increase when stimuli from two or more talkers are presented within a block as compared to when stimuli from only one talker are presented within a block. That is, the difficulty will be compounded if subjects must try to adapt to differences between talkers.

Given Yamada and Tohkura's [6] "range effects", we should not observe large changes in the overall rate of "R" response if we manipulate the range of cues to /r/ and /l/ by varying the amount of talker variability -- subjects should adapt such that they respond "R" 50% of the time. However, if subjects set criteria based on the overall range of cues they hear within a

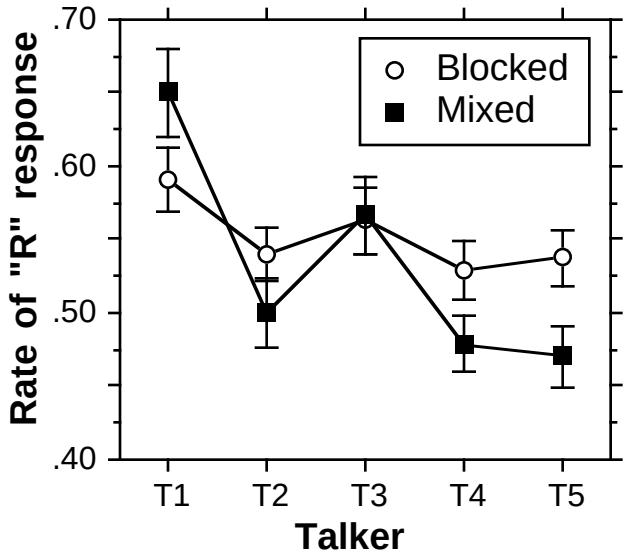

Figure 1: The interaction of talker and talker condition observed in Experiment 1.

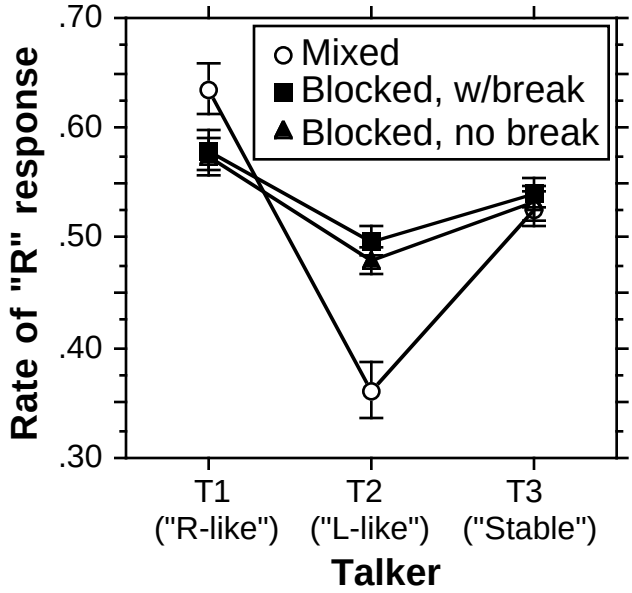

Figure 2: The interaction of talker and talker condition observed in Experiment 2.

block of trials, those criteria should become less successful as the number of talkers increases, and errors will increase.

In the next section, we will review perceptual results in which this sort of talker by talker condition interaction was found. Then, we will turn to new analyses of acoustic correlates to those perceptual results.

2. EXPERIMENTS

The data we will examine here comes from two previous experiments [2]. In the experiments, we considered the question of whether Japanese listeners are able to adapt to talker-specific differences in cues to /r/ and /l/, or if added variability due to talker differences instead influences session-specific criteria, as did the “range” manipulations in [6].

In Experiment 1, the task was as follows. Subjects were aurally presented with a real English word spoken by a native speaker of AE. The word had /r/ or /l/ in initial position. A minimal pair of English words that differed only by the presence of /r/ or /l/ (e.g., "right" and "light") were orthographically presented to subjects on a CRT. The aurally presented word was one of the pair, and the subjects' task was to choose the orthographic word they thought matched the aural presentation. The subjects were 27 college-age native speakers of Japanese. All subjects performed this task in two talker conditions. In the *blocked* condition, the stimuli were presented in five blocks. The stimuli in each block were produced by a different talker, but the stimuli within a block were produced only by one talker. In the *mixed* condition, the stimuli were also presented in five blocks, but in each block, equal numbers of stimuli were produced by each of the five talkers and mixed in random order. Rate of "R" response was used as a measure of bias. "R"-rate is simply the proportion of "R" responses, and is equivalent to "yes"-rate measures of bias used in some varieties of detection theory.

The overall "R"-rate was close to .50 in every block. However, in the mixed-talker condition, "R"-rate to particular talkers differed substantially from "R"-rate in the blocked condition, leading to the significant interaction of talker and talker condition ($F(4,104) = 6.11$, $p < .001$) shown in Figure 1.

Experiment 2 was very similar to Experiment 1. The task was the same, and the design differed mainly in two respects. First, only three talkers were used: T1 (who appeared "R-like" in Experiment 1, in that "R"-rate to his stimuli increased in the *mixed* condition relative to the *blocked* condition), T2 (who appeared "L"-like), and T3 (who appeared "stable," as "R"-rate to his productions did not vary substantially between talker conditions). Second, there were two variations of the *blocked* condition: the *no-break* condition was identical to the *blocked* condition in Experiment 1, but in the *break* condition, subjects listened to piano music for two minutes between each talker block. This manipulation was used to examine whether subjects were "tuning" to talkers in the blocked condition and whether that tuning persisted over a break (see [2] for details). The subjects were 34 college-age native speakers of Japanese.

The results of Experiment 2 were similar to those of Experiment 1. As in Experiment 1, we found a significant interaction of talker and talker condition ($F(4,132) = 17.34$, $p < .001$). As can be seen in Figure 2, the interaction of talker condition and talker are due to the rate of “R” response being much lower to the “L-like” talker than to the others in the mixed condition.

Our interpretation of these *talker* by *talker condition* interactions is that our subjects are applying one strategy in each talker condition: they attempt to divide the range of cues to /r/ and /l/ within a block into two categories. In the *blocked* condition, this is indistinguishable from adjusting categories to adapt to differences between talkers, as only one talker is heard in each block. In the *mixed* condition, however, this strategy leads to errors, since subjects are not familiar enough with how /r/-/l/ cues can vary from talker to talker to adapt on-

line to talker changes. Instead, they form categories based on the range of cues *across* talkers.

When would subjects adopt such a strategy? We predicted that subjects could be divided into three distinct groups, only one of which would use this strategy. Subjects with low accuracy (near chance), would not have well-enough formed categories for /r/-/l/ to be able to use such a strategy. Subjects with high accuracy (near native speaker levels) would have well-enough formed categories that they would not need to resort to strategies. The subjects in between should be primarily responsible for the talker by talker condition interaction. It can be seen from Figure 3 that the prediction held: "R"-rate in the two talker conditions in Experiment 2 is shown with subjects divided into groups based on average accuracy (we observed similar patterns in Experiment 1).

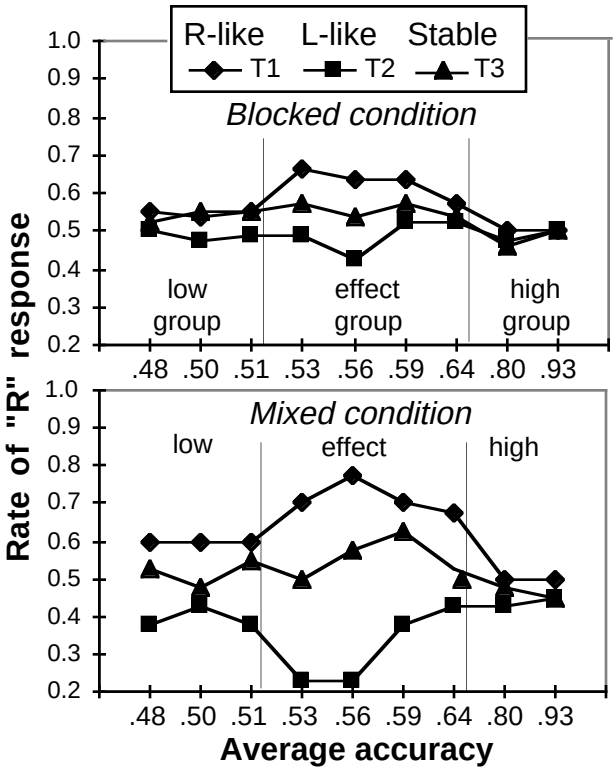

Figure 3: Rate of "R" response to each talker by average accuracy in *blocked* and *mixed* talker conditions in Experiment 2. The *blocked* condition is the average of the *break* and *no-break* conditions. Each point represents the average rate of "R" response for four subjects, with the exception of the points for subjects with average accuracy of .93, which are the averages for two subjects. The numbers on the X axis are labels only.

3. ACOUSTIC ANALYSES

Given our interpretation of the perceptual results (i.e., that subjects are sorting stimuli across talkers in the mixed condition based on ranges of sometimes invalid cues to /r/-/l/ identity), an important question is what cues subjects are attending to. We identified a number of possible cues, examined their ecological validity, and their correlation with subjects' rate of "R" response for each stimulus in the two experiments. The cues we examined were initial center-frequencies of F1, F2, F3 in the /r/ or /l/ portion of the stimulus, and the duration of the /r/ or /l/ portion of a stimulus. For the stimuli used in Experiment 1, there was one ecologically-valid cue: F3 was completely reliable. The maximum measured value of F3 for /r/ stimuli was 1980 Hz (M = 1638 Hz, range = 1031 Hz to 1980 Hz) and the minimum measured value of F3 for /l/ stimuli was 2097 Hz (M = 2826 Hz, range = 2097 Hz to 3326 Hz). A simple regression was computed for the relation of consonant identity to F3: r^2 = .853. The next highest match was with F2: r^2 = .122. F2 is not a good cue for "R-L" decisions. Although the mean measured values were 1163 Hz for /l/ stimuli and 1040 Hz for /r/ stimuli, the ranges were similar: 753 Hz to 1610 Hz for /l/ stimuli, and 660 Hz to 1475 Hz for /r/ stimuli. F1 and duration were not reliable cues (r^2 = .0002 and .001, respectively). For the subset of the stimuli used in Experiment 2, we found similar degrees of reliability for the cues we examined.

What cues did subjects use? We must note that we cannot determine whether subjects ever used F3 in our experiments: because of the division between /r/ and /l/ stimuli in F3, high accuracy will always be correlated with F3, even if subjects are relying on another cue we have not identified (but see Yamada and Tohkura [6], who were able to analyze sensitivity to F3 by systematically changing it in synthetic stimuli).

Our first examination of the correlations between cues and the "R"-rate to particular stimuli in Experiment 1 indicated that there was little change in the correlations between blocked and mixed talker conditions. A subsequent examination of the correlation of cue values and "R"-rate across talkers within each accuracy group (low, effect, high) indicated a stronger reliance on duration by subjects in the effect groups (and some subjects in the other groups) in both experiments (with stimuli with relatively longer initial steady state durations being identified as "R"). Tables 1 and 2 show the change in fit F2, F3 and duration between blocked and mixed talker conditions for each accuracy group in the two experiments (F1 is not included because r^2 was never greater than .08). As can be seen in the tables, the only cue for which there were consistent, substantial increases in correlation with "R"-rate in the mixed talker condition was duration (especially for the effect groups).

4. DISCUSSION

The change in correlation between "R"-rate and duration between blocked and mixed talker conditions (especially among the "effect" group subjects) provides further support for our hypothesis that our subjects were attempting to find session specific criteria for /r/-/l/ decisions. In the mixed-talker condition, the correlation of talker-specific "R"-rates and duration of the initial /r/ or /l/ portion of the stimuli increased substantially (relative to the blocked-talker condition). This suggests that subjects may have chosen duration as a cue that could be applied across talkers.

The new analysis of subject differences allowed us to identify groups of subjects with different strategies for /r/-/l/ decisions. The knowledge that subjects with relatively low average accuracy that is somewhat above chance adopt session-specific criteria for /r/-/l/ identification may prove useful for non-native

contrast training (see [9] for some preliminary studies of the effects of talker variability on /r/-/l/ training).

These results demonstrate that it is possible to identify strategies used by non-native speakers who do not have well-formed categories for non-native speech contrasts. It is important to note that the effect is not specific to the /r/-/l/ contrast and Japanese speakers: we have found similar talker by talker condition interactions when AE and Japanese speakers are asked to distinguish the Hindi dental and retroflex stop contrast [8]. Theoretically, this common result indicates the importance of considering the effects of talker variability in non-native speech perception. Practically, the implications for adult non-native language learning are that effective training may require talker variability (indeed, there is evidence that in some cases training with only one talker may fail [1]) and that how talker variability is structured (e.g., mixed or blocked) may also be an important consideration (see [9], in which we compare blocked- and mixed-talker training).

5. REFERENCES

1. Lively, S. E., Logan, J. S., & 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 perceptual categories. *J. Acoust. Soc. Amer.*, 44, 180-186.

2. Magnuson, J.S., & Yamada, R.A. (1994). Talker variability and the identification of American English /r/ and /l/ by Japanese listeners. *J. Acoust. Soc. Amer.*, 95 , 2872.

3. Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of vowels. *J. Acoust. Soc. Amer.*, 24, 175-184.

4. Yamada, R.A. (1993). Effect of extended training on /r/ and /l/ identification by native speakers of Japanese. *J. Acoust. Soc. Amer.*, 93, 2391.

5. Yamada, R.A., & Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. *Proc. 1990 Internat'l Conf. on Spoken Language Processing*, 757-760.

6. Yamada, R.A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, 52, 376-392.

7. O'Connor, J.D., Gerstman, L.J., Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1957). Acoustic cues for the perception of initial /w, r, l/ in English. *Word*, 13, 25-43.

8. Magnuson, J.S., Pruitt, J.S., & Yamada, R.A. (in preparation). The effects of talker variability on the perception of the Hindi dental vs. retroflex stop contrast by native speakers of English and Japanese.

9. Magnuson, J.S., & Yamada, R.A. (1995). *Proc. 1995 Internat'l Congress of Phonetic Sciences*, 306-309.

Group	Cue	Blocked r^2	Mixed r^2	Change in fit: mixed-blocked
All	F2	**.154**	**.184**	.030
	F3	**.559**	**.534**	-.025
	Duration	.020	**.111**	**.091**
Effect	F2	**.190**	**.271**	**.081**
	F3	**.371**	**.362**	-.009
	Duration	.053	**.237**	**.184**
High	F2	**.161**	**.199**	.038
	F3	**.787**	**.805**	.018
	Duration	.001	.010	.009
Low	F2	.001	.006	.005
	F3	.022	.002	-.020
	Duration	.005	.018	.013

Table 1: Results of simple regressions of cue values to rate of "R" response for each accuracy group in Experiment 1. Values of r^2 greater than .100 are presented in boldface. Change in fit (r^2) is presented in the right-most column, and changes greater than .050 are presented in boldface.

Group	Cue	Break r^2	No break r^2	Mixed r^2	Change in fit: mixed - break	Change in fit: mixed - no break
All	F2	.058	.077	**.113**	**.055**	.036
	F3	**.311**	**.365**	**.311**	0	**-.054**
	Duration	.036	.054	**.209**	**.173**	**.155**
Effect	F2	.080	.095	**.122**	.042	.027
	F3	**.190**	**.218**	**.177**	-.013	-.041
	Duration	.087	.089	**.339**	**.252**	**.250**
High	F2	.027	.041	.036	.009	-.005
	F3	**.732**	**.753**	**.719**	-.013	-.034
	Duration	.005	.001	0	-.005	-.001
Low	F2	.003	.009	.033	.030	.024
	F3	.000	.008	.003	.003	-.005
	Duration	.016	.045	**.115**	**.099**	**.070**

Table 2: Results of simple regressions of cue values to rate of "R" response for each accuracy group in Experiment 2. Values of r^2 greater than .100 are presented in boldface. Change in fit is presented in the right-most column, and changes greater than .050 are presented in boldface.

ACKNOWLEDGMENTS

We thank Prof. David Pisoni for providing us with the stimuli, Dr. Yoh'ichi Tohkura for suggesting the analysis of subject differences, and Prof. Winifred Strange, Prof. Kevin Munhall, Prof. Howard Nusbaum, Dr. Hideki Kawahara, Inge-Marie Eigsti and Dr. John S. Pruitt for comments which substantially improved this research.