DO CROSS-LANGUAGE SPECIALIZATIONS IN PHONETIC PERCEPTION EXTEND ACROSS NOVEL ACOUSTIC TRANSFORMS?

Anita Wagner, Paul Iverson & Stuart Rosen

Department of Speech Hearing and Phonetic Sciences, UCL, UK
a.wagner@ucl.ac.uk; p.iverson@ucl.ac.uk; s.rosen@ucl.ac.uk

ABSTRACT

Native Japanese speakers can have difficulty learning the English /r/-/l/ distinction, and they likewise are poorer than native English speakers at discriminating acoustic differences at the /r/-/l/ identification boundary. This study investigated the specificity of this specialization by comparing the discrimination of /r/ and /l/ under stimulus transforms that disrupted the natural acoustics of the stimuli but still allowed them to be identified as /r/ and /l/ (i.e., sinewave speech, vocoded speech). The results demonstrated that the cross-language discrimination difference held across these stimulus transforms. This suggests that language specialization is speech specific rather than occurring in general auditory processing, given that there is little pure-auditory reason that these stimuli should all be processed similarly.

Keywords: cross language, speech perception

1. INTRODUCTION

Individuals become specialized through development to perceive native-language (L1) phonetic contrasts, such that second-language (L2) contrasts can become difficult for adults to learn. For example, Japanese listeners have notorious difficultly in distinguishing English /r/ and /l/ [1, 3, 4, 8].

The present study investigated to what extent this specialization occurs based on auditory processing of these phonemes or higher-level categorization processes. Early work suggested that specialization occurs at the level of linguistic categorization [8]. That is, Japanese listeners were shown to have poorer discrimination sensitivity for /r/ and /l/ than did English speakers, but Japanese listeners were shown to discriminate /r-l/ F3 components virtually native like when these acoustic cues were presented in isolation (perceived as non-speech). This suggested that Japanese listeners have the auditory acuity to distinguish /r/ and /l/, and only have difficulty with these acoustic contrasts when phonemic categories are applied.

More recent work has suggested, however, that lower-level processes may also become specialized during language learning. For example, Mandarin tones can be discriminated more accurately by Mandarin than English speakers, but Mandarin speakers also have an advantage for analogous non-speech pitch contours, even at early stages of central auditory processing [6]. Behavioral research on other phonetic contrasts have likewise found that phonemic categorization does not always predict discrimination abilities, suggesting that a lack of perceptual sensitivity for L2 phonemes can have an origin in pre-categorical auditory/phonetic processing [4, 10].

One difficulty with auditory accounts of speech perception is that speech can remain intelligible even when the acoustic form of speech is radically disrupted. For example, sinewave speech (replacing formants and fricatives with frequency-modulated sinusoids; [9]) and vocoded speech (amplitude-modulated frequency bands with sinusoidal or noise carriers; [11]) can both be intelligible even though neither have the pitch or voice quality of normal speech. This demonstrates that listeners can recognize speech by perceiving spectral-temporal patterns that do not preserve traditional acoustic features, and indicates that surface-level auditory properties may not be critical to speech recognition.

In the present study, we investigated whether such manipulations of acoustic form have an impact on /r/-/l/ discrimination differences for Japanese and English adults, in order to assess the extent to which this specialization is based on lower-level auditory processing of these speech signals. We created five synthetic speech continua: a synthetic /r-l/ continuum modeled on natural speech, three vocoded speech continua with different carriers (sinusoid, noise, and harmonic),

and sinewave speech. All continua were designed so that the endpoints would be identifiable as /r/ and /l/ by L1 English speakers, but they had very different acoustic forms. L1 speakers of English and Japanese were asked to identify endpoints of the continua as /r/ or /l/, and then discriminated stimuli that crossed the /r/-/l/ identification boundary. The aim was to determine whether the cross-language discrimination difference for /r/ and /l/ was or was not preserved when the acoustic form was disrupted.

2. METHOD

2.1. Subjects

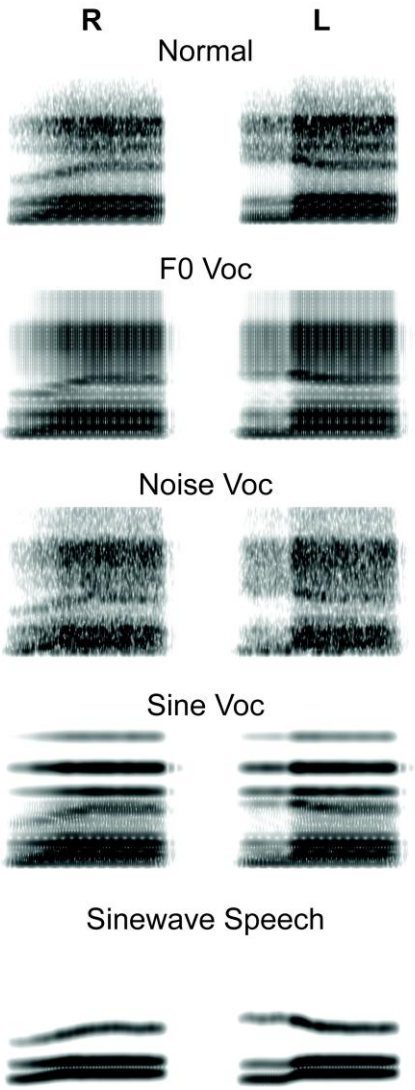
Eight native southern British English speakers and 8 native Japanese speakers were tested. The Japanese listeners were screened using an /r/-/l/ identification task with natural stimuli, to ensure that all participants had /r/-/l/ difficulty. The ages ranged from 18-24 for the British speakers, and from 20-42 for Japanese listeners. On average, the Japanese listeners started learning English at the age of 13. None of the participants reported any hearing problems.

2.2. Stimuli

All acoustic transforms were applied to a /ra/-/la/ stimulus continuum based upon best exemplars found in a previous study [3] that were generated using a Klatt synthesizer [5]. Respectively for /r/ and /l/, F3 varied from 2403 to 3508 Hz, the duration of the initial closure (i.e., before the transition to the vowel) varied from 31 to 96 ms, and the duration of the formant transition varied from 81 to 16 ms. The continua were equally divided into 76 steps. See Figure 1 for example spectrograms.

The vocoded conditions were created using MATLAB. The center frequencies of the channels were selected so that they would be equally spaced with regard to the basilar membrane [2], but with a closer spacing of channels in the critical F3 region. The center frequencies of the twenty channels were: 137, 229, 348, 502, 704, 966, 1307, 1751, 2086, 2221, 2365, 2518, 2680, 2851, 3033, 3226, 3431, 4059, 5334, and 6992 Hz. The amplitude envelopes of these bands were low-pass filtered at 300 Hz. The amplitude envelopes within each channel were multiplied with different carriers to create the three continua: F0 Voc, Noise Voc, and Sine Voc. In the F0 Voc condition the carrier consisted of a 220-Hz pulse train. In the Noise Voc condition the carrier was random noise. In the Sine Voc condition the amplitude envelope was multiplied with a sinusoid at the band center frequency.

Figure 1: Spectrograms of /r/-/l/ endpoints for the five stimulus conditions.



A fifth condition was sinewave speech [9] created using Praat. For this series, the frequency values and amplitude envelopes for the first three formant transitions were extracted from the synthetic stimuli by means of LPC analysis. For each stimulus, three sinusoids corresponding to the F1, F2 and F3 frequency and amplitude values were generated, and then added up to create a 3-formant sinewave continuum.

2.3. Procedure

Listeners began the experiment with a training task that was designed to familiarize them with each

acoustic transform; they saw a written sentence on the computer screen and interactively matched this with an acoustically transformed version of this spoken sentence. Listeners then performed a screening task in which they gave forced-choice /r/-/l/ identification judgments on initial-position minimal pair words recorded by multiple speakers, under each of the acoustic transforms.

The main task was a discrimination experiment that assessed the ability of subjects to discern acoustic differences that straddled the /r/-/l/ boundary under each stimulus condition. Listeners heard three stimuli on each trial (two same and one different) and had to choose the one that they thought was different. An adaptive procedure was used such that listeners began by discriminating the endpoint stimuli, and the acoustic differences between the stimuli became smaller or larger depending on whether their responses were right or wrong, such that the procedure converged on the acoustic difference that yielded 71% correct responses [7].

3. RESULTS

Figure 2 displays the average identification results for /r/- and /l/-initial words. As expected, Japanese listeners' identification was near chance under all conditions, whereas English speakers were nearly uniformly 100% correct.

Figure 2: Boxplots of identification accuracy for each of the five stimulus conditions for English (black) and Japanese (white) listeners. Boxplots display the quartile ranges of scores.

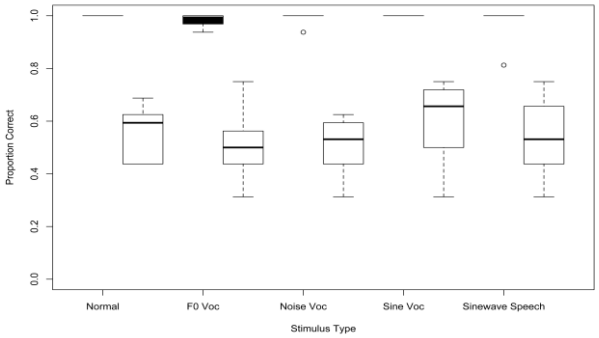
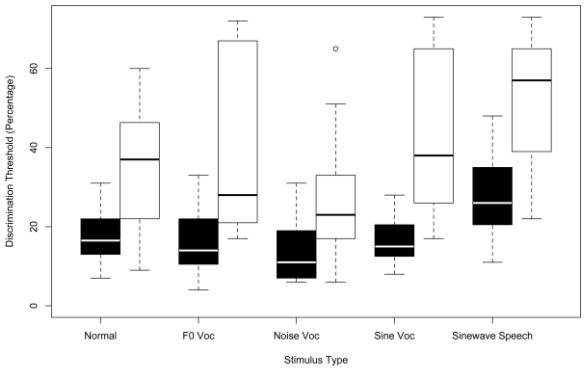


Figure 3 displays boxplots of the discrimination thresholds for synthetic stimuli that straddled the English /r/-/l/ boundary. The threshold is expressed as percentage of the stimulus range (e.g., 100% would mean that the threshold was at the stimulus endpoints), and a lower threshold indicates greater auditory sensitivity. As expected, English speakers had lower discrimination thresholds near the /r/-/l/ boundary when the stimuli sounded the most like natural /r/ and /l/ productions (i.e., the "normal" condition), but this cross-language difference appeared to extend to highly unnatural transformations of the acoustic form, such as sinewave speech. This was confirmed in an ANOVA analysis. That is, there was a significant main effect of language background, $F(1,14) = 31.4$, $p < 0.001$, indicating a cross-language difference. There was also a main effect of condition, $F(4,14) = 13.3$, $p < 0.001$, indicating that there was some variation in the thresholds due to the acoustic transform (e.g., higher thresholds in the sinewave speech condition). However, there was no significant interaction, $p > 0.05$.

Figure 3: Boxplots of discrimination thresholds for each of the five stimulus conditions for English (black) and Japanese (white) listeners. Boxplots display the quartile ranges of scores.



4. DISCUSSION

The results of this study thus demonstrate that the cross-language difference in discrimination of /r/ and /l/ by Japanese and English speakers extend to intelligible stimuli that have highly unnatural acoustic forms. Previous work had suggested that these kinds of cross-language differences may have an origin in auditory processing for speech. For example, the patterns of auditory sensitivity that Japanese speakers have for the /r/-/l/ contrast cannot be explained by top-down categorization [4]. Electrophysiological evidence has suggested that the /r/-/l/ differences emerge earlier, both in terms of anatomy and processing time, than lexical processing potentials [12]. Also, work with other types of contrasts (Chinese tone) have found differences in non-speech pitch processing [6]. However, the present results suggest that the cross-language differences for /r/ and /l/ cannot be based on processes that are tuned to the surface-level

acoustics of speech, because disruptions of the natural acoustic form have no effect on this difference.

A straightforward interpretation of this result is that the cross-language difference between Japanese and English speakers for /r/ and /l/ may have its origin in phonological categorization. That is, all of the stimuli could be reliably categorized as /r/ and /l/ by native English listeners, and were near chance in identification accuracy by Japanese listeners. There were likewise cross-language differences in discrimination for all continua despite their varying acoustic similarity to normal speech. The results may thus be in accord with the original claims of Miyawaki, et al. [8], who found discrimination differences for /r/-/l/ but not for isolated F3 transitions that sounded like whistles rather than speech.

It is possible, however, that the specialization could occur at a level in between auditory processing and phonological categorization, such as in pre-categorical phonetic processing. Although all stimulus transforms disrupted the auditory surface forms, all conditions preserved the more abstract spectral-temporal patterning of acoustic information that is characteristic of speech. This suggests that the cross-language differences found here are speech specific, given that there is no purely auditory reason for why these stimuli ought to have been discriminated the same. However, it is plausible that phonetic processing could also operate based on this more abstract spectral-temporal information, and that differences could emerge prior to the stimuli being categorized as /r/ and /l/. Additional research is required to fully explore whether the identification of stimuli as /r/ and /l/ is necessary for this cross-language discrimination difference, or if it is possible to find cross-language differences for stimuli that fall short of sounding like /r/ and /l/.

5. ACKNOWLEDGEMENTS

This research was funded by a grant from the Wellcome Trust.

6. REFERENCES

[1] Goto, H. 1971. Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia* 9, 317-323.

[2] Greenwood, D.D. 1990. A cochlear frequency-position function for several species--29 years later. *Journal of the Acoustical Society of America* 87, 2592-2605.

[3] Hattori, K., Iverson, P. 2009. English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *Journal of the Acoustical Society of America* 125, 469-479.

[4] Iverson, P., Kuhl, P.K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., Siebert, C. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47-B57.

[5] Klatt, D.H., Klatt, L.C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America* 87, 820-857.

[6] Krishnan, A., Gandour, J.T. 2009. The role of the auditory brainstem in processing linguistically relevant pitch patterns. *Brain and Language* 110, 135-148.

[7] Levitt, H. 1971. Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America* 49, 467-477.

[8] Miyawaki, K., Strange, W., Verbrugge, R.R., Liberman, A.M., Jenkins, J.J., Fujimura, O. 1975. An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics* 18, 331-340.

[9] Remez, R.E., Pardo, J.S., Piorkowski, R.L., Rubin, P.E. 2001. On the bistability of sinewave analogs of speech. *Psychological Science* 12, 24-29.

[10] Schouten, B., Gerrits, E., van Hessen, A. 2003. The end of categorical perception as we know it. *Speech Communication* 41, 71-80.

[11] Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M. 1995. Speech recognition with primarily temporal cues. *Science* 270, 303-304.

[12] Zhang, Y., Kuhl, P.K., Imada, T., Iverson, P., Pruitt, J., Stevens, E.B., Kawakatsu, M., Tohkura, Y., Nemoto, I. 2009. Neural signatures of phonetic learning in adulthood: A Magnetoencephalography (MEG) study. *Neuroimage* 46, 226-240.