FREE CLASSIFICATION PREDICTS DISCRIMINATION OF FINNISH VOWELS BY NAÏVE JAPANESE LISTENERS

Danielle Daidone[1], Ryan Lidster[1], and Franziska Kruger[2]

[1]University of North Carolina Wilmington and [2]Indiana University
daidoned@uncw.edu, lidsterr@uncw.edu, fkruger@indiana.edu

ABSTRACT

This study explores whether listeners' groupings of vowels by similarity in a free classification task can be used to predict the discriminability of those vowels. Recent research has shown that free classification results correlate strongly with discrimination, which could potentially enable researchers to predict discrimination across a variety of phenomena not easily studied with other existing tools. However, the method's generalizability has yet to be tested. In particular, we are interested in the utility of free classification for listeners with a small first language vowel system.

Naïve Japanese listeners, whose native system contains five vowels, performed a free classification task and an oddity discrimination task with the eight vowels of Finnish. Results showed that free classification was a strong predictor of discrimination scores. These findings suggest that free classification is a useful method for predicting the discriminability of stimuli, even when listeners have a small number of native categories.

Keywords: free classification, discrimination, non-native perception, oddity, laboratory phonology

1. INTRODUCTION

Free classification (FC), also known as free sort, is a task in which listeners group stimuli simply according to how similar they perceive them to be. Participants are presented with all of the stimuli at once on the screen, and they listen to them as many times as they like and in any order that they choose. There are no researcher-imposed labels or categories; rather, the participants form groups of any size, arranging together those stimuli that they believe to sound similar. This method has been used for dialect similarity and accentedness research [1, 3, 5] and has recently been extended into investigating the similarity of non-native segments [6]. There is emerging evidence that a free classification task can be used to predict the discriminability of non-native contrasts as well [7]. In that study, we found that FC was a strong predictor of discrimination scores in experiments examining the perception of German vowels and Finnish phonemic length by American English listeners.

FC has many potential advantages as a task type: completing the task does not require knowledge of orthography or any linguistic terminology, and since all stimuli are available for comparison at once, FC typically takes a short time. Since FC can in principle be used for any type of stimuli, it could potentially enable researchers to predict confusability and differences across a variety of phenomena not easily studied with other existing tools. However, FC's generalizability has yet to be tested, and the particular phonological system of listeners' first languages (L1) may make free classification more or less predictive due to the way results are scored. When analyzing FC results, each pair of stimuli are coded dichotomously: 1 if the pair is grouped together by that participant, and 0 if not. To obtain more continuous ratings of perceived similarity, researchers average the results of multiple trials for stimuli of the same category (e.g., presenting participants with the same vowel in 2-3 different voices or consonant contexts) and/or by averaging across many participants [see 4]. For example, if Participant X grouped both male and female /e/ with female /ɛ/, but put the male /ɛ/ token in a different group, the /e-ɛ/ contrast overall would have a grouping rate of 50% for that individual (two possible pairings made, and two not made). Similarly, if Participant X grouped all of the /e/ and /ɛ/ tokens together, but Participant Y put all /e/ and /ɛ/ tokens in separate groups, the combined grouping rate for /e-ɛ/ would be 50%.

Given this type of analysis, one concern for generalizability is the potential for participants to form large groups of stimuli that they consider to be vaguely "similar" but are nonetheless discriminable. In such a scenario, FC grouping rates would fail to predict differences for discriminability within the large groups. This is especially a concern when listeners have few L1 categories; thus, many non-native stimuli could represent within-category differences for them, making it more likely for them to group many stimuli together regardless of their phonetic differences.

Therefore, to test whether FC is still a useful predictor of discrimination accuracy in the case where the listeners' L1 has few categories, we tested Japanese listeners on their perception of Finnish short

vowels. Japanese has only five vowels /a e i o ɯ/ which can be long or short, raising concern that Japanese listeners may not make sufficiently detailed groupings of Finnish vowels to predict differences in their perception performance across contrasts.

2. METHOD

2.1. Participants

Forty-one L1 Japanese listeners participated. Results from 14 participants were later excluded for various reasons: 10 had timeouts on more than 5% of oddity trials, one had lived abroad in Austria, one did not follow instructions on the FC task, one had had speech therapy, and one had phonetics training. All 27 participants (age range 18-31, M=20) whose results were included passed a bilateral hearing screening and did not report any prior hearing or speech problems. All had studied English and had one semester of study in a third language as part of university requirements (9 Chinese, 6 German, 6 French, 3 Russian, 1 Korean), in which all participants reported having beginner-level proficiency. Because none had prior exposure to Finnish, we have termed the participants "naïve" listeners with respect to Finnish.

2.2. Stimulus material and vowel analysis

The 8 short vowels of Finnish /i e y ø u o æ ɑ/ were embedded in alveolar (/tVhVt/) and velar contexts (/kVhVt/). In each of the stimuli, the same vowel was repeated, e.g. /tihit/; /kuhuk/. Three female native speakers of Finnish from Helsinki recorded the stimuli. The recordings were then analyzed in Praat [2]. Vowel boundaries were marked at a zero-crossing at the beginning and end of each vowel and the values for f0, F1, F2 and F3 were extracted for each speaker's tokens at 25%, 50%, and 75% of the total vowel duration. This was done for both vowels within a word. Following Flynn [9], formants were normalized with Gerstman's formula. The average normalized F1 and F2 values at the vowel midpoint for each context separately are shown in Figure 1. A square represents values from Speaker 1, a diamond for Speaker 2, and a triangle for Speaker 3. Solid shapes represent the alveolar context; border-only shapes represent the velar context.

2.3. Procedure

Participants completed a hearing screening, free classification (FC) task, oddity task, background questionnaire, and a perceptual assimilation task (not discussed here). All participants wore headphones and completed the tasks in this order.

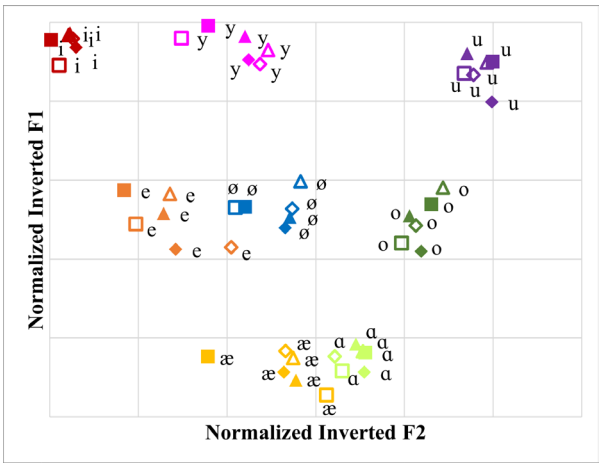

Figure 1: Average formant values of the Finnish stimuli.

FC was administered via PowerPoint. On each slide, a 16x16 grid was shown on the left and sound files in randomly numbered squares were presented on the right. Oral and written instructions in Japanese asked participants to make groups of any size according to what sounds they thought were similar to each other. Participants were instructed to click on the sound files to listen to them and drag them onto the grid to form groups. They could listen to each sound file and rearrange groups as many times as they liked. Groups had to consist of at least two sound files, and they were instructed to ignore differences in speaker. Participants completed two slides, one for each consonant context. Each slide presented 24 sound files (8 vowels x 1 context x 3 speakers), and the order of slides was counterbalanced across participants. Participants took approximately 10-20 minutes to complete the task.

The oddity task was administered via a web browser using jsPsych [8]. During each trial, participants saw three different colored robots, one representing each Finnish speaker, and an X on the screen. Each robot "said" a word, and participants clicked on the one that said something different, or clicked X to indicate that all three robots said the same thing. There were first 8 practice trials using words with a /o-e/ contrast (e.g. /tehet-tohot/) to confirm participants understood the task. If they made more than one mistake on the practice, they repeated it until they could pass and move on to the experimental trials.

The experiment tested discrimination performance on 10 different vowel contrasts: /u-y/, /u-ø/, /y-ø/, /o-ø/, /e-ø/, /e-i/, /e-æ/, /æ-ɑ/, /ɑ-o/, and /ɑ-i/, chosen because they were thought to present a range of difficulties based on pilot data from American English listeners. Each target contrast appeared once in the 6 possible sequences of "different" trials (ABB, BAA, ABA, BAB, AAB, BBA) and twice in each of the 2 possible "same" trials (AAA, BBB), per

consonant context, giving a total of 200 trials ([6 different trials + 4 same trials] * 10 contrasts * 2 consonant contexts). Trials had an ISI of 400ms, ITI of 1000ms, and a time limit of 3500ms to respond. The trials were blocked by consonant context with a short break in between, presented in random order within blocks. The task took approximately 25 minutes to complete.

3. ANALYSIS AND RESULTS

3.1. Free Classification

The researchers coded which stimuli were grouped together by each participant on each slide. An R script was then used to tabulate all possible pairs of stimuli, coded as 1 if participants put them in the same group and 0 if not. For example, if participants grouped the /ø æ ɑ/ tokens of one speaker together, then the pairs /ø-æ/, /ø-ɑ/, and /æ-ɑ/ were all given a 1 for that participant. The total number of times each pairing was made was divided by the maximum number possible, resulting in the percent of the time those stimuli were rated as similar by participants. These formed a matrix of similarity ratings between each pair of stimuli. Overall similarity ratings are displayed in Table 1. Cell shading reflects the degree of similarity.

Table 1: Free classification similarity ratings.

	ɑ	æ	e	i	o	ø	u	y
ɑ		0.70	0.04	0.01	0.13	0.35	0.07	0.01
æ			0.08	0.01	0.07	0.35	0.02	0.00
e				0.07	0.01	0.07	0.01	0.01
i					0.00	0.00	0.01	0.05
o						0.14	0.76	0.01
ø							0.10	0.19
u								0.02

Similarity ratings were then inverted (1-similarity) to create dissimilarity matrices which were analyzed using SPSS 28's ALSCAL function for multidimensional scaling (MDS) with a convergence criterion of 0.001 [4]. MDS is an analytical procedure used to analyze the similarity of multiple stimuli simultaneously by finding a positioning for all stimuli in one common space that best recreates the "distances" (dissimilarities) observed in the data. After examining solutions ranging from 1 to 5 dimensions, the 3-dimensional solution was found to be the best fit for the data for both consonant contexts separately based on the low stress values (/t/: 0.111, /k/: 0.099), high R^2 (0.937 for both), and location below the elbow of the stress plots [4]. Only minute differences between consonant contexts were found, so for brevity, Figures 2a and 2b display the averaged locations across both consonant contexts in 3D perceptual space as estimated by MDS. Both the grouping rates as well as the averaged distances between the rescaled stimuli locations in 3D MDS space were compared to oddity results to evaluate their predictive power.

The MDS solution visually shows how acoustics are warped in perceptual space, suggesting that /u/ and /o/ were perceived as very similar to each other, while /e/ and /ø/ were not. The overlap between /u/ and /o/ is a surprising result that differs from the pilot data with American English listeners.

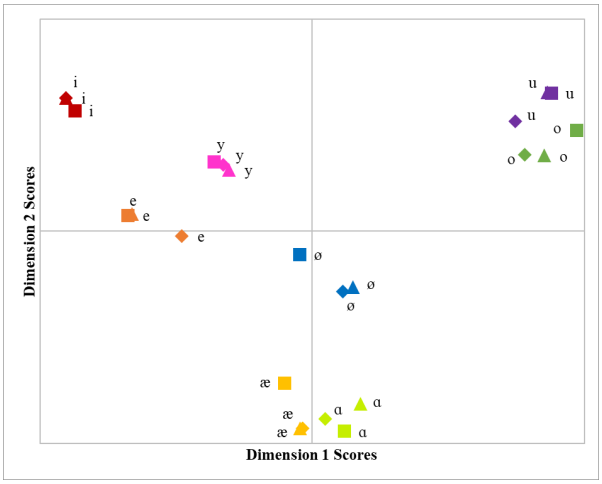

Figure 2a: MDS solution for Dim1 x Dim2.

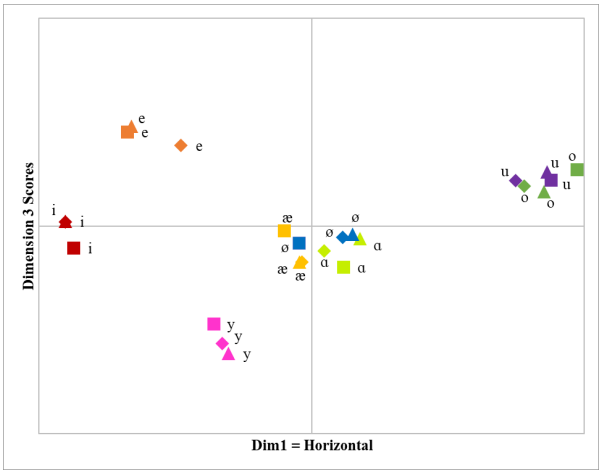

Figure 2b: MDS solution for Dim1 x Dim3. (The same data as Figure 2a if viewed "from above")

To help interpret these results, we compared dimension scores to our measurements of the vowels' acoustic properties, and also coded vowels for two phonological properties: rounding (1 for rounded, 0 for unrounded), and specifically front-rounding (1 for front rounded vowels, 0 for all else), since they represent a category of vowel not present in Japanese at all. The results, shown in Table 2, indicate that Japanese listeners seemed to predominantly use vowel backness (F2) and height (F1) to distinguish stimuli. Dimension 3 scores are less easily

interpretable, but it appears that Japanese listeners are separating front-rounded vowels from others. Duration and pitch were not related to how Finnish short vowels were grouped on the FC task.

Table 2: Correlations with MDS solutions by context.

	F1	F2	F3	Dur	f0	Δf0	Rnded	Front-Rnded
Dim1	0.25	**-0.95***	-0.33	0.09	0.25	-0.16	0.56*	-0.11
Dim2	**-0.79***	0.08	0.33	0.24	0.22	-0.05	0.30	-0.16
Dim3	-0.44*	-0.03	-0.28	0.10	-0.01	0.05	0.56*	**0.70***

* $p < 0.05$ after False Discovery Rate corrections.

3.2. Oddity

Oddity performance on each contrast was evaluated for accuracy. Accuracy rates by contrast are displayed in Figure 3. Mauchly's test for Sphericity was significant, so we ran a one-way ANOVA with sphericity corrections, showing that accuracy rates significantly differed by contrast ($F_{(5.4,140)} = 41.669$, $p < 0.001$, $\eta^2 = 0.499$). Post-hoc pairwise comparisons with Bonferroni corrections revealed 25 significant comparisons, including that /æ-ɑ/ was more difficult than every other contrast. The next most difficult contrasts all involved /ø/, showing that the Japanese listeners found it confusable with several other Finnish vowels, whereas /u-y/ was relatively easy for Japanese listeners.

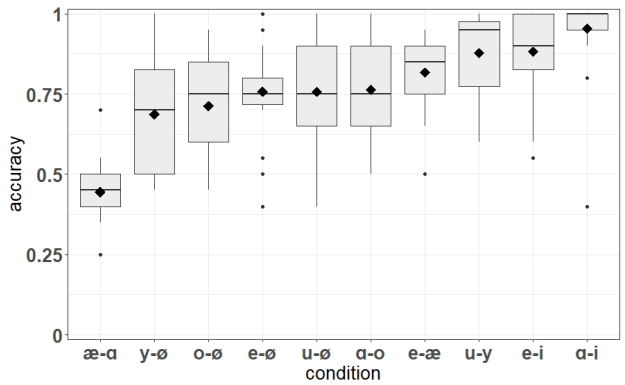

Figure 2b: Accuracy on oddity task by contrast.

3.3. Task Comparison

Linear regressions were run to determine how well the FC results predicted the accuracy of discrimination of contrasts in oddity. For the first regression analysis, the independent variable was FC similarity rates by contrast and the dependent variable was oddity accuracy scores by contrast. This regression equation was significant, $F_{(1, 18)} = 90.54$, $p < .001$, $r = 0.913$, $R^2 = .834$ (CI[.72, .93]), showing that FC similarity rates strongly predicted performance on oddity. In the second regression, distances between stimuli in the 3D MDS solution were entered as the independent variable, also yielding a significant result, $F_{(1, 18)} = 28.93$, $p < .001$, $r = 0.785$, $R^2 = .616$ (CI[.05, .91]).

4. DISCUSSION AND CONCLUSION

The performance of Japanese listeners on a variety of Finnish vowel contrasts ranged widely by contrast in ways that may not have been easily predicted simply by examining the phonemic inventories alone. For example, /u-y/ was surprisingly easy, and /y/, /u/, /o/, and /e/ all appear to be similarly confusable with /ø/, while /æ-ɑ/ was quite difficult to discriminate for Japanese listeners.

We tested whether an FC task could be used to predict this pattern of performance, and the results show that FC grouping rates were a strong predictor of variable discrimination. For example, /æ-ɑ/ was found to largely overlap according to FC, while /u/ and /y/ were perceived as much more distinct from each other. This shows that, despite Japanese having relatively few native language vowel categories, listeners nevertheless demonstrated sensitivity to phonetic detail in their grouping behavior. The surprisingly tight correlation between FC grouping rates and oddity accuracy ($r = 0.913$) suggests that the task is a valuable tool for investigating non-native perception.

Furthermore, by examining the FC results using Multidimensional Scaling (MDS), we obtained an informative visualization of perceptual warping. FC also provided a concise means of predicting performance on contrasts beyond those tested in this current experiment, opening questions for future research. For example, the unexpected overlap between /u/ and /o/ suggests that Japanese listeners may have considerable difficulty discriminating this contrast. If true, determining why this is different from Japanese listeners' attested ease with English /u-o/ [10], for example, could help advance our understanding of Japanese perception in general.

The task features of FC—in particular the lack of researcher-imposed labels and the ability to present all stimuli together—make it a promising tool for phonologists investigating the perception of a wide variety of phenomena. More research is needed to determine whether this predictive power generalizes further to suprasegmentals and other features. By continuing to explore the relationship between task performance across languages, we believe phonologists can arrive at a more explanatory theory of L2 perception.

5. ACKNOWLEDGEMENTS

We thank Aaron Albin, Rebecca Muth, and the IU L2 Psycholinguistics Lab for their contributions.

6. REFERENCES

[1] Atagi, E., Bent, T. 2013. Auditory free classification of nonnative speech. *Journal of Phonetics 41*, 509-519.

[2] Boersma, P., Weenink, D. 2015. Praat: doing phonetics by computer. Version 5.4.08 http://www.praat.org/.

[3] Bradlow, A. R., Clopper, C., Smiljanic, R., Walter, M. A. 2010. A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication 52*, 930–942.

[4] Clopper, C. G. 2008. Auditory free classification: Methods and analysis. *Behavior Research Methods 40*, 575–581.

[5] Clopper, C. G., Pisoni, D. B. 2007. Free classification of regional dialects of American English. *Journal of Phonetics 35*, 421–438.

[6] Daidone, D., Kruger, F., Lidster, R. 2015. Perceptual assimilation and free classification of German vowels by American English listeners. *Proc. 18th ICPhS.* University of Glasgow.

[7] Daidone, D., Lidster, R., Kruger, F. 2023. Free classification as a method for predicting perceptual discriminability of non-native contrasts. *Studies in Second Language Acquisition*, 1-27.

[8] de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods 47*, 1–12.

[9] Flynn, N. 2011. Comparing vowel formant normalisation procedures. *York Papers in Linguistics Series 211*, 1–28.

[10] Nishi, K., Kewley-Port, D. 2007. Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research 50,* 1496–1509.