Variation of [v] in Cook Islands Māori

Quartz Colvin[Rutgers University]

Variation of [v] in Cook Islands Māori

1.0 Introduction

In this paper, we will do a statistical analysis of [v] across a sample of islands in Cook Islands Māori (CIM). It's known that in many dialects and other varieties of Māori, this phoneme can be realized as [w] or [v]. This paper aims to take a statistical approach to this generalization.

This paper has four sub-questions to investigate. First, how does w~v duration vary by island and second, how does intensity for these phonemes vary by island. The other two questions are about identifying information about the surface forms of w~v by island. Specifically, we will model f0 and f2 to determine whether certain islands have a voiced phoneme realized. Finally, the f2 model will help us determine which islands have higher rates of [w]s surfacing and which have more [v]s surfacing.

1.1 Background (CIM)

Cook Islands Māori is an Eastern Polynesian language classified as *Vulnerable*. It is very closely related to Aotearoa Māori, but is definitely different from it.

It's known across both Aotearoa (NZ) Māori and CIM that some speakers regularly pronounce [v] as [w], but it isn't clear to anyone who does this more and what conditions it.

2.0 Methods

3.0 Data

This data set contains a lot of information, so we will need to tidy it and only keep the relevant data for the analysis. The following code chunk shows how the data was tidied and what was kept from the untidy data set. I don't show a preview of the untidy data set here because it had 17 columns and most of the columns will be taken out. In addition, it includes some metadata that I did not ask permission to share.

```
dat <- untidy_data |>
  drop_na() |>   # remove rows that have NA values
  filter(TextGridLabel == "v") |>   # filter by 'v' for analysis
  pivot_longer(      # deal with all the f0 columns
    cols = c("f0_20.point",
             "f0_50.point",
             "f0_80.point"),
    names_to = "percent",    # names to 'percent' column
    values_to = "hz"      # values to one column called 'hz'
  ) |>
  arrange(island, Word) |>
  select(speaker, island,   # group by 'island', then 'word'    # pick which columns I
         word = Word,
         percent, hz, f2 = F2_midpoint, duration,
         intensity = intensity_midpoint) |>
  write_csv("tidy_cim.csv")    # put tidy data in new .csv
```

In the new data set, we need to fix the names that we've converted to values under 'percent'. This is shown in the following code chunk.

```
# fixing the original column names to only be the percent
# since they are all values for f0

dat$percent[dat$percent == 'f0_20.point'] <- "20"

dat$percent[dat$percent == 'f0_50.point'] <- "50"

dat$percent[dat$percent == 'f0_80.point'] <- "80"
```

Now that we have tidied this data set, here is a preview showing what it looks like.

speaker	island	word	percent	hz	f2	duration	intensity
TA	Atiu	ava	20	128.5916	964.3925	0.08	54.08202
TA	Atiu	ava	50	127.0908	964.3925	0.08	54.08202
TA	Atiu	ava	80	127.4540	964.3925	0.08	54.08202
TA	Atiu	ava	20	160.6464	950.2333	0.07	58.06061
TA	Atiu	ava	50	157.4397	950.2333	0.07	58.06061
TA	Atiu	ava	80	158.2820	950.2333	0.07	58.06061

There are many more rows than this, but if all rows were included, it would be too long of a table to show here. To explain the column names, "Percent" refers to the 20% 50% and 80% marks of the phoneme's total duration and "HZ" refers to the f0 measurement (in Hz) at that mark. "f2" is formant 2 measured in Hz at the midpoint of the phoneme. "Duration" is the total duration for the phoneme and "Intensity" is the intensity extracted from the midpoint of the phoneme.

4.0 Analysis

This section readdresses the main questions about the w~v alternation. First, how does duration differ by island and second, how does intensity vary by island? Finally, do any islands have regular occurrence of [w] instead of [v] surfacing. For this final question, we will compare f0 and f2 separately.

Note that in all of my models, I did not control for where w~v occurs in the word, since I am looking at general frequency information and not doing a phonological analysis of specific environments in which [w] or [v] occurs more.

4.1 Duration by island

First, we compare the duration of [v] across the four islands. To do this, I will fit a *linear model* below to track the relationship between the duration of w~v and which island a speaker is from. For this model, duration is the dependent variable and island is the independent or predictor variable.

Before we fit the model, we need to see if fitting a model for duration is appropriate. The visual qq-plot test for duration shows that we need to transform the values for duration in order to fit a model.

To normalize the duration variable, we need to log-transform the data. So, we make a new column where we apply 'log()' to the 'duration' column with 0.1 added to it, and then add the new column to the dataset. The following boxplot shows the normalized durations by island.

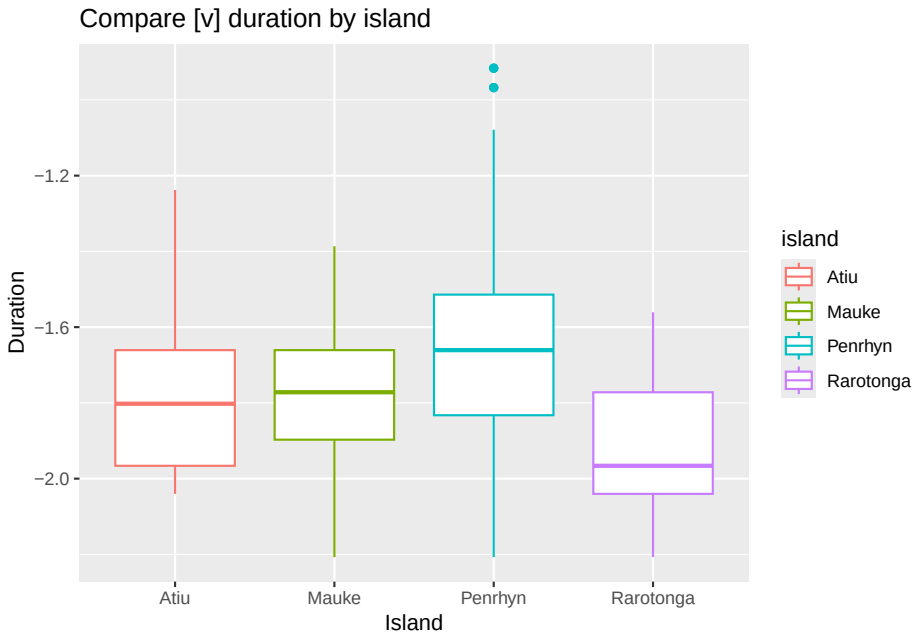

Figure 1

A few things are visually clear from this comparative boxplot. Most obviously,

Penrhyn had the largest range of w~v durations and Rarotonga had the smallest range. Rarotonga also had the lowest average duration for w~v and Penrhyn had the highest average duration. Interestingly, Mauke and Rarotonga didn't seem to have outliers in w~v durations, but Atiu and Penrhyn had a few outlier data points.

4.2 Intensity by island

Next, in the same fashion we did to compare duration of [v] tokens, we can compare the intensity midpoints for the tokens across the islands. First we need to look at a qq-plot to see if fitting a model for intensity is an appropriate decision.

The qq-plot test showed that intensity *is* normally distributed, so we can fit a model for intensity. Once again I fit a linear model for intensity (the dependent variable) by island (the predictor variable). The results are shown via a boxplot again.

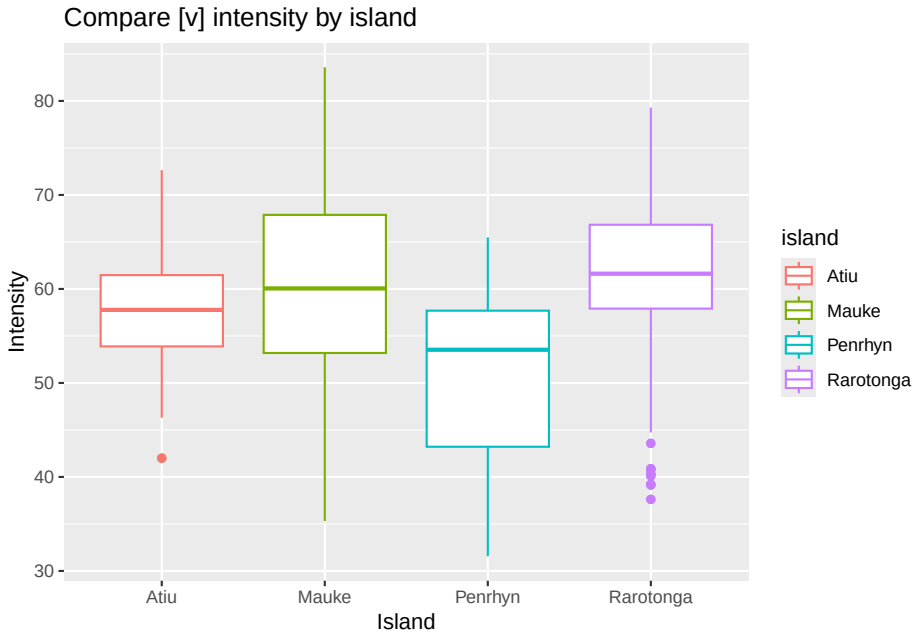

Figure 2

Looking at the comparative boxplot, we can see how intensity midpoints vary across these four islands. The average intensity values were similar for all four, but Penrhyn had

the lowest mean. These averages varied by less than 10, so they are quite similar. Mauke had the highest range for intensity midpoints, while Atiu had the smallest. Rarotonga had the most outliers and the other three islands had basically no outliers.

4.3 Voicing by island

Now we can investigate how voicing of w~v varies by island. The f0 values used for this were extracted at the 20%, 50% and 80% points of the duration of the target phoneme.

Again, we first need to check if fitting a model for this variable is appropriate. If we do another visual qq-plot test, we find that 'Hz' is normally distributed and thus we can fit a model.

This time, I will show the model results in a scatterplot with a facet for each island. These plots show how F0 (measured in Hz) changes over the duration of the w~v tokens.

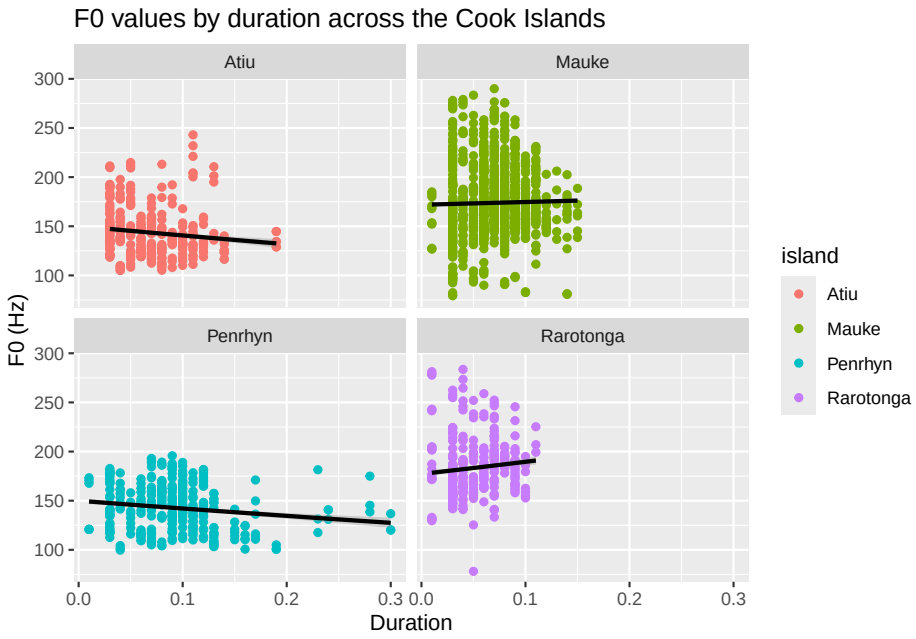

Figure 3

As noted earlier, duration clearly differs between these four islands. The slopes for Atiu and Penrhyn show that over the duration of the phoneme, f0 decreases. This slope isn't

steep, but it is present. Mauke and Rarotonga have upwards slopes, showing that over the duration of [v], the f0 tends to increase. Rarotongan Māori has the sharpest increase in Hz over the duration of the phoneme.

Mauke is the only one that seems to *regularly* have voiceless [w] in place of 'v', but it also has the biggest overall range for f0 Hz. The other three islands typically have f0 values $>$ 100 Hz, so these tokens are voiced.

In summary, it seems that f0 was not a good predictor for the [w] vs. [v] question, but it did reveal that Mauke's [w] phoneme has a higher rate of voicelessness when it is realized.

4.4 w~v distribution by island

Finally, we can revisit the main question of this article: which islands have more [w] or more [v]? One of the primary acoustic differences between [w] and [v] is that [w] has a lower f2. Here, we define "low f2" as <850 Hz. To investigate this question, we will fit a model for f2 by island.

If we do a visual qq-plot test, we find that the raw data for f2 are *not* normally distributed. Before we fit the model, we need to use 'log()' to transform the data and put it in a new column like we did for duration.

Since we normalized the f2 values in the model, we also have to transform the cutoff point of 850 Hz by the same log transformation we did for the real f2 values. If we do this, the cutoff point in our model is 6.75.

Visually, all of these four islands have instances of [w], and their means are all slightly (or much) higher than the cut-off point. Additionally, all of the islands except Penrhyn have *many* tokens where 'v' surfaces as [w].

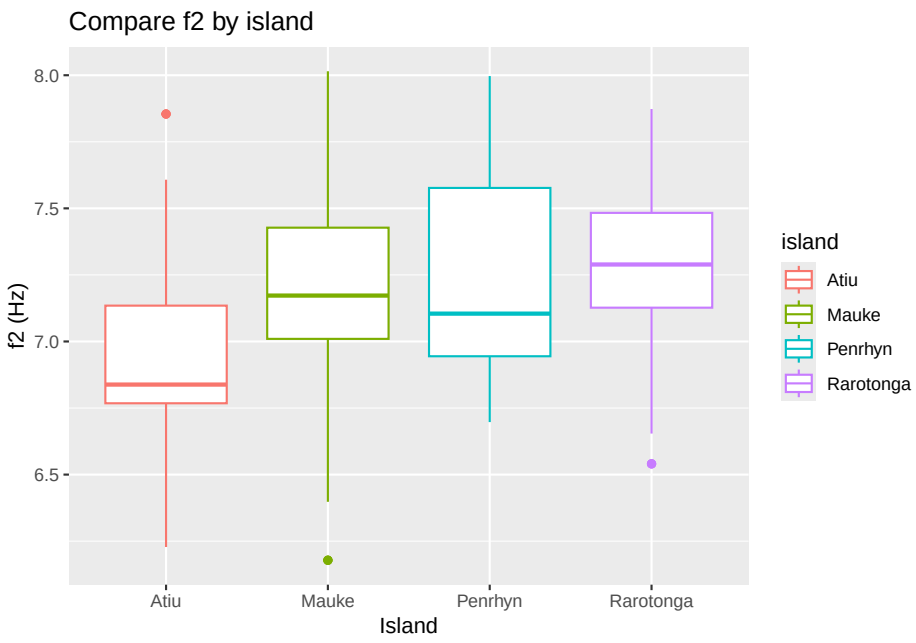

Figure 4

5.0 Formal results

To interpret the results of each model, we will do t-tests for the normalized versions of each model. If $p < 0.05$ then the difference is not random chance. For the t values, a larger t value indicates a greater difference between the means.

For *normalized duration*, t=-408.11 and p-value < 2.2e-16, and there is a 95% confidence interval. For *intensity*, t=286.21, p-value < 2.2e-16 and there is a 95% confidence interval. For *hz* (f0), t=207.57, the p-value is < 2.2e-16, and we have a 95% confidence interval. For *normalized f2*, t=1060.9, p-value is < 2.2e-16 and we have a 95% confidence interval.

All of these p-values are < 0.05, so duration, intensity, f0 (Hz column), and f2 are all statistically significant interactions with island. T values for duration and f2 were the furthest from 0, so these acoustic properties varied the most depending on which island a speaker was from.

I now want to pose the question of what percent of w~v surface as [w] and [v] for each island in the data, now that we have the values and a cutoff value for f2. This requires another bit of data transformation to calculate percentages. The first transformation step is adding a new column that says "v" if the value in log_f2 is greater than the cutoff value, and outputs "w" if the value in the 'log_f2' column is lower than the cutoff value.

island	Atiu	Atiu	Mauke	Mauke	Penrhyn	Penrhyn	Rarotonga	Rarotonga
surface_form	v	w	v	w	v	w	v	w
freq	264	54	1005	42	327	6	303	6

I manually calculated the percentages of [w]s surfacing in each group (island). The percent of [w]s in Atiu was 16.98%, 4.01% for Mauke, 1.80% for Penrhyn and 1.94% for Rarotonga.

We can also transform the data in a similar way to calculate the percentages of [w] that are voiced in each island.

island	Atiu	Mauke	Mauke	Penrhyn	Penrhyn	Rarotonga	Rarotonga
voiceless_w	318	1023	24	332	1	308	1

The percentage of voiceless [w]s in Mauke was 2.29%, for Penrhyn it was 0.30% and for Rarotonga it was 0.32%.

5.1 Data analysis

I used R (Version 4.5.0; R Core Team, 2024) and the R-packages *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *ds4ling* (Version 1.0; Casillas, 2025), *forcats* (Version 1.0.0; Wickham, 2023a), *ggplot2* (Version 3.5.2; Wickham, 2016), *ggpubr* (Version 0.6.0; Kassambara, 2023), *here* (Version 1.0.1; Müller, 2020), *lme4* (Version 1.1.37; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *Matrix* (Version 1.7.3; Bates, Maechler, & Jagan, 2025), *papaja* (Version 0.1.3; Aust & Barth, 2024), *purrr*

(Version 1.0.4; Wickham & Henry, 2025), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidyr* (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019) and *tinylabels* (Version 0.2.5; Barth, 2023) for all my analyses.

6.0 Conclusion

In conclusion, this paper found that duration of w~v, intensity, f0, and f2 are all correlated with what island a CIM speaker is from. The strongest correlations with island are duration and f2. The island with the longest average duration for w~v is Penrhyn and the shortest average is Rarotonga. The island with the highest average intensity midpoint is Rarotonga, and the shortest is Penrhyn.

F2 data showed us that all islands have some small percent of [w]s surfacing instead of [v], but Atiu has the highest rate of realizing [w]s instead of [v]s. Furthermore, when w~v surfaces as [w], formant data from f0 revealed that only Mauke has multiple instances of it surfacing voiceless, at 2.29% of [w]s being voiceless.

References

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://doi.org/10.32614/CRAN.package.papaja

Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Maechler, M., & Jagan, M. (2025). *Matrix: Sparse and dense matrix classes and methods*. https://doi.org/10.32614/CRAN.package.Matrix

Casillas, J. (2025). *ds4ling: Datasets and functions developed for SPAN658: Data science for*

linguists. Retrieved from https://github.com/jvcasillas/ds4ling

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from https://www.jstatsoft.org/v40/i03/

Kassambara, A. (2023). *Ggpubr: 'ggplot2' based publication ready plots*. https://doi.org/10.32614/CRAN.package.ggpubr

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Müller, K. (2020). *Here: A simpler way to find your files*. https://doi.org/10.32614/CRAN.package.here

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. https://doi.org/10.32614/CRAN.package.tibble

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. https://doi.org/10.32614/CRAN.package.forcats

Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*. https://doi.org/10.32614/CRAN.package.stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://doi.org/10.32614/CRAN.package.dplyr

Wickham, H., & Henry, L. (2025). *Purrr: Functional programming tools*.

https://doi.org/10.32614/CRAN.package.purrr

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data.* https://doi.org/10.32614/CRAN.package.readr

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data.* https://doi.org/10.32614/CRAN.package.tidyr