

Variation of [v] in Cook Islands Māori

Quartz Colvin

1 Introduction

In this paper, we will do a statistical analysis of [v] across a sample of islands in Cook Islands Māori (CIM). It is 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 the duration of w~v vary by island and second, how does intensity for w~v 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 realize a devoiced phoneme. Finally, an 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 indigenous to the Cook Islands. CIM is classified as *Endangered* and has little intergenerational transmission. It is very closely related to Aotearoa Māori (New Zealand Māori), but is definitely different enough to be classified as a separate, but related language.

It is known across both Aotearoa Māori and CIM speakers that some speakers regularly pronounce their [v]s as [w]s, but it is not clear to anyone who does this more and what conditions it. In hopes of answering this question, I will compare various aspects of w~v phonemes across a sample of the Cook Islands. The islands in this study are Atiu, Mauke, Penrhyn, and Rarotonga.



Figure 1: Map of Cook Islands.

In the map above, you can see that Penrhyn is far in the northern group while the other three islands in our sample are part of the southern group. Lastly, it's relevant to note that Rarotongan CIM is considered to be the standard dialect of CIM and it hosts the capital of the Cook Islands.

2 Methods

The data for this project comes from a **large digital corpus** managed by **PARADISEC** (Nicholas 2012). Much of the data in this corpus was collected for Nicholas's dissertation work and while there is a *lot* of data, it is largely unorganized.

In order to make large corpora like this as useful for linguistic research or revitalization projects as possible, the data needs to be fully annotated and organized. Given the sheer amount of data, it would take humans too long to annotate it all, and thus a team of linguists and Cook Islanders have started working together to develop NLP tools to accelerate this documentation (Coto-Solano et al. 2018).

As part of this broader collaborative NLP project, they have trained an improved forced alignment tool for CIM. So far, forced-aligned data has been used to investigate the glottal stop phoneme (Coto-Solano et al. 2018), but this segmented data can be used to do all kinds of phonetic analysis. The data in this paper comes from the PARADESEC corpus and was aligned by phoneme, and then corrected. The final Praat TextGrids are also available in the corpus. All acoustic measurements for these phonemes were extracted from Praat.

To reiterate, all of the data collection and processing to a usable format for this project was not done by me. Full credit for all of those previous steps goes to the linguists who worked on data collection and processing the data (Coto-Solano et al. 2018).

3 Data

The following large data set has many columns. Here I show a preview of the data set and then discuss which columns will be relevant to the analysis later in the paper.

```
## Rows: 2,513
## Columns: 17
## $ speaker      <chr> "AAN", "AAN", "AAN", "AAN", "AAN", "AAN", "AAN", "A...
## $ island       <chr> "Rarotonga", "Rarotonga", "Rarotonga", "Rarotonga", ...
## $ Filename     <chr> "AAN-RRAT8-018", "AAN-RRAT8-018", "AAN-RRAT8-018...
## $ TextGridLabel <chr> "a", "v", "e", "v", "a", "a", "v", "a", "v", "a", "...
## $ Word         <chr> "rave", "rave", "rave", "va'ine", "va'ine", "rava", ...
## $ PreviousLabel <chr> "r", "a", "v", NA, "v", "r", "a", "v", NA, "v", "r"...
## $ FollowingLabel <chr> "v", "e", NA, "a", "v", "v", "a", NA, "a", "v"...
## $ start        <dbl> 1.10, 1.13, 1.16, 1.73, 1.76, 0.75, 0.80, 0.87, 0.6...
## $ end          <dbl> 1.13, 1.16, 1.19, 1.76, 1.79, 0.80, 0.87, 0.90, 0.6...
## $ duration     <dbl> 0.03, 0.03, 0.03, 0.03, 0.03, 0.05, 0.07, 0.03, 0.0...
## $ f0_20.point  <dbl> 144.0532, 141.7901, 143.1055, 263.7962, 265.8672, 2...
## $ f0_50.point  <dbl> 143.0606, 142.2725, 142.6605, 269.8593, 261.0071, 2...
## $ f0_80.point  <dbl> 141.8219, 143.3506, 142.2108, 270.8541, 263.7023, 2...
## $ F1_midpoint  <dbl> 356.4547, 335.4101, 341.6010, 537.7380, 584.4075, 6...
## $ F2_midpoint  <dbl> 1755.6225, 1999.8371, 2362.3051, 1040.5499, 1419.93...
## $ F3_midpoint  <dbl> 2797.351, 3192.717, 3280.530, 2898.961, 2876.166, 2...
## $ intensity_midpoint <dbl> 40.29265, 39.22013, 40.47651, 52.75762, 55.72719, 5...
```

Figure 2:

First, note that there is no 'id' column but there is a column for the speaker's initials. I will use speaker in place of 'id' because that is what was in this original data set. The 'island' column states what island that speaker is from. This data only contains data from the four islands Atiu, Mauke, Penrhyn and Rarotonga. Duration corresponds to the total duration of the phoneme, and intensity is measured at the midpoint of the phoneme.

F0 values were taken at 20%, 50% and 80% of each phoneme's total duration. Later, we will opt to use only the F0 values at 50%, but that will be discussed more later. Finally, F2_midpoint is f2 measured in Hz at the midpoint of that phoneme. All of the other columns are either straight-forward information or will be irrelevant for this paper.

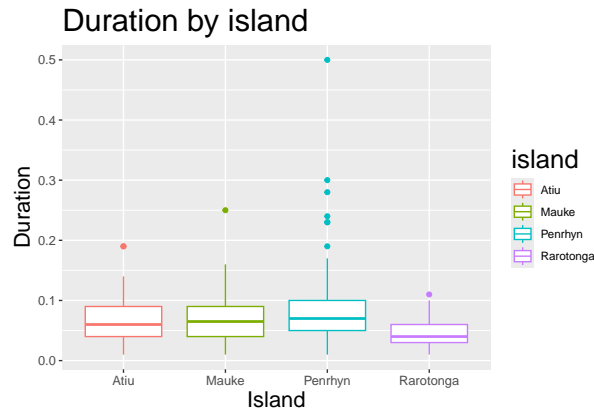


Figure 3: Duration boxplot

4 Analysis

This section readdresses the main questions about the $w \sim v$ 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 $w \sim v$ 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. Another important thing to note is that for each island, there were different numbers of speakers contributing to the data set. All of the Atiu data came from one speaker and it was the same for Penrhyn. The Mauke data came from four speakers and the Rarotonga data came from five speakers. Since there are different numbers of speakers from each of these islands, the models I will use to investigate each question will be linear mixed models.

4.1 Duration by island

The first step of comparing duration of $w \sim v$ across these four islands is to make a smaller, tidy data set that only contains the relevant information. Below, I filter the data so that it only contains rows with the “v” phoneme. Then I arrange it by island and then by word, and I select only the columns “speaker”, “island”, “word” and duration. Finally, I save this filtered data set as a new csv file. After we have done this tidying process, the duration data set looks like this.

Table 1: Duration tidy data.

speaker	island	word	duration
TA	Atiu	ava	0.08
TA	Atiu	ava	0.07
TA	Atiu	ava	0.14
TA	Atiu	ava	0.11
TA	Atiu	ava	0.05
TA	Atiu	ava	0.13

For a visual comparison, we will look at boxplots comparing duration across these four islands.

The boxplot shows that average (mean) durations were pretty similar across these three islands. But for some reason, Penrhyn had a much larger range of durations than the other islands.

Now that we have an idea of what the general differences and similarities are for duration, we can fit the linear mixed effect model. In the code chunk, you can see that I account for the fact that number of speakers varies across the islands in the data set, and also that there are many instances of each word.

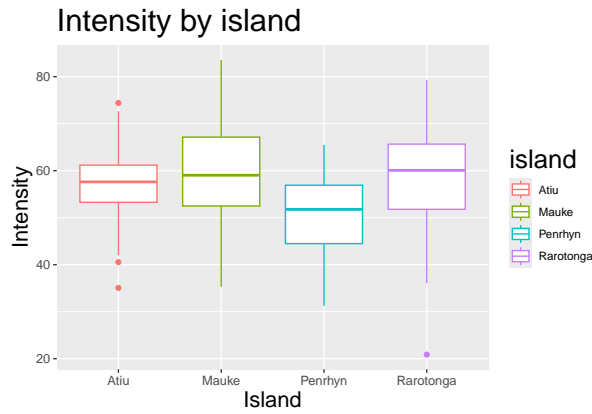


Figure 4: Intensity boxplot

These instances of each ‘word’ are repeated by each speaker multiple times and by several speakers multiple times.

```
mod_dur <- lmer(duration ~ island + (1|speaker) + (1|word), data = tidy_dur)
```

4.2 Intensity by island

Next is the intensity comparison across the four islands. Starting from the raw data set, we must filter it again so that it only has instances of the “v” phoneme and we filter everything except the “speaker”, “island”, “word”, and “intensity” columns. Once the new data set has been made, it looks roughly like the following table, but extended of course.

Table 2: Duration tidy data.

speaker	island	word	intensity
TA	Atiu	ava	54.08202
TA	Atiu	ava	58.06061
TA	Atiu	ava	48.40481
TA	Atiu	ava	54.18315
TA	Atiu	ava	49.59224
TA	Atiu	ava	53.07335

In the same way we did for the duration question, we will first look at a boxplot to visually see any clear differences in intensity across the Cook Islands. The code chunk to make the plot is below, and then the actual plot is printed below it.

Initially, it is clear that Penrhyn has the lowest mean intensity for w~v. Mauke and Rarotonga have nearly the same mean intensity, and Atiu’s mean is slightly lower, but not as low as Penrhyn’s mean. Now that we know the general patterns, we can fit our linear mixed effects model. Due to the same reasons as before, I model speaker and word as mixed effects in the intensity by island model. The code chunk below shows how I fit the model in R.

```
mod_intense <- lmer(intensity ~ island + (1|speaker) + (1|word), data = tidy_intense)
```

4.3 Voicing by island

Moving on to the formant questions, first we will investigate how voicing of w~v differs across these four islands. In acoustic measurements, an F0 lower than 80 Hz is considered voiceless.

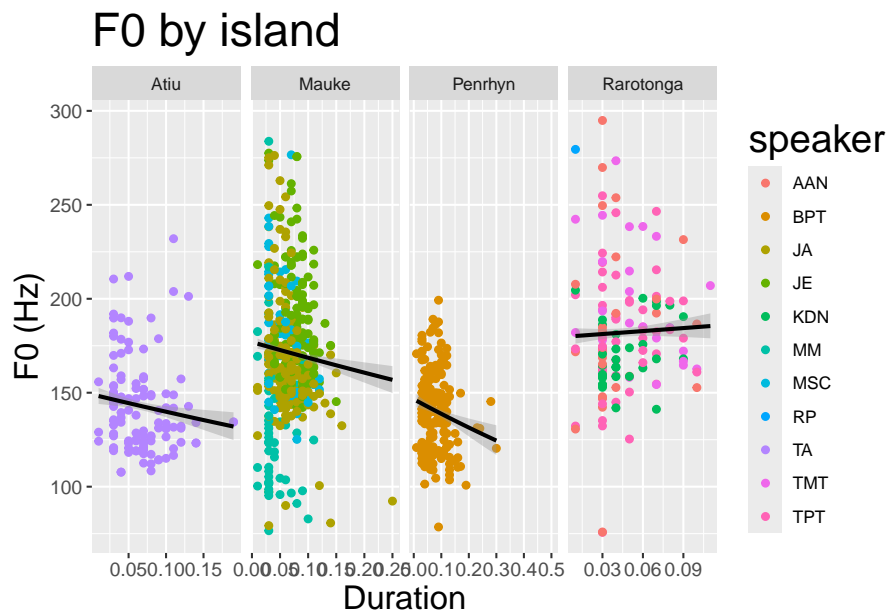


Figure 5: F0 by island

This time, we tidy the raw data set in a similar way but with one additional step. Since we have f0 values at three different intervals of the same morpheme, if we were to put all of those formant values in a new column and the percentage in another new column, then it would fabricate 3 times as many tokens that were not elicited from the speakers.

In order to limit this autocorrelation issue, I chose to only keep f0 values at the 50% interval of the phoneme. I chose the 50% interval because the middle of the duration seemed like the smartest choice to represent the average f0 value across the phoneme's duration. The code chunk below shows how I tidied this mini-data-set to answer the voicing question.

Once we have made this new, tidy data set, it looks like this. Again, if I showed the full table, it would have more than this one word and this one island. This is just the top few rows of the sorted table.

Table 3: F0 tidy data

speaker	island	word	f0	duration
TA	Atiu	ava	127.0908	0.08
TA	Atiu	ava	157.4397	0.07
TA	Atiu	ava	134.1155	0.14
TA	Atiu	ava	141.8125	0.11
TA	Atiu	ava	118.2318	0.05
TA	Atiu	ava	142.7506	0.13

This time, to compare f0 across the islands visually, we can plot this data onto a scatterplot. As you can see, I created an individual facet for each island and each color represents data from a particular speaker.

It's clear that the most drastic slope is in the Penrhyn plot, and the data for the islands with more speakers are much more spread out across the plot. To formally compare these different variables, we can again use a linear mixed effects model to see how f0 varies across these four Cook Islands. The model fitting is shown in the code chunk below.

```
mod_f0 <- lmer(f0 ~ island + (1|speaker) + (1|word), data = tidy_f0)
```

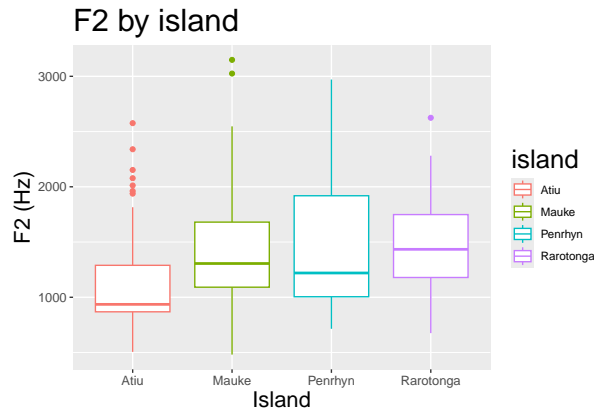


Figure 6: F2 boxplot

4.4 $w \sim v$ distribution by island

Finally we turn to the main research question: is there a correlation between island and how many $[w]$ tokens surface? The acoustic measurement indicating a $[w]$ is **less than 850 Hz**. For this question, we tidy the raw data set once again so that it only has data for “v”, speaker, island, word, and f2 values. Once the data has been tidied, it generally looks like this.

Table 4: F2 tidy data.

speaker	island	word	f2
TA	Atiu	ava	964.3925
TA	Atiu	ava	950.2333
TA	Atiu	ava	2575.9621
TA	Atiu	ava	961.5736
TA	Atiu	ava	984.4541
TA	Atiu	ava	896.0441

To visually compare the f2 values across these islands, we generate another boxplot.

From looking at the f2 boxplot, it’s clear that Atiu has the lowest mean f2. The other three islands had more similar means, but Mauke and Penrhyn had some instances of $w \sim v$ that had extremely high F2 measurements. Regardless, the boxes for each island look different, so we know what to expect when we fit a linear mixed effects model. The code chunk for fitting this model is below.

```
mod_f2 <- lmer(f2 ~ island + (1|speaker) + (1|word), data = tidy_f2)
```

5 Results

This section interprets the statistical results for each individual model, and interprets a nested model comparison for each of the four models from the previous section.

5.1 Duration

Data were analyzed using a linear mixed effects model (as implemented by the lme4 (version 1.1-37) package in R 4.5.0 (2025-04-11 ucrt)). This model included *duration* as the dependent variable, and *island*, *speaker* and *word* as fixed factors.

The fixed effects for this individual model are in Table 5. The **standard error (SE)** for each island were very small, around 0.01, indicating that the mean of the sample is an accurate reflection of the

actual populations mean. The **t-values** for each island except Atiu were all lower than 2, indicating that they are not statistically significant between the islands, unless a speaker is from Atiu. The **degrees of freedom (df)** for Mauke and Penrhyn were a little under 6 and the df for Rarotonga was a bit above 6. For each island, the **betas** are very close to 0, which indicates there is not a linear relationship between duration and island.

	betas	SE	t-val	df
Atiu	0.064761	0.013130	4.932	6.339137
Mauke	-0.001702	0.014390	-0.118	5.858906
Penrhyn	0.019477	0.018068	1.078	5.690570
Rarotonga	-0.018439	0.014557	-1.267	6.292463

Table 5: Duration fixed effects

As for the random effects for the duration model, these are shown in Table 6. The **standard deviations** for both word and speaker were very small, indicating that the data for these two variables were very clustered around the mean. **Variance** was also extremely small.

group	name	Std. Dev.	Variance
word	(intercept)	0.02707	0.0007328
speaker	(intercept)	0.01231	0.0001517
Residual		0.03192	0.0010189

Table 6: Duration random effects

Nested model comparison.

The model summary for the nested model comparison is in Table 7. For this, I utilized the ‘anova()’ function in R and omitted the columns I don’t interpret. Significance of main effects and all possible interactions were assessed using hierarchical partitioning of the variance via nested model comparisons. Experiment-wise, alpha was set at 0.05. The model summary table for the nested model comparison for duration is shown in the following table.

	AIC	BIC	logLik	Chisq	Pr(>Chisq)
dur_null	-3287.577	-3277.981	1645.789	NA	NA
dur_island	-3367.804	-3343.815	1688.902	86.22701	0
dur_add	-3397.532	-3368.745	1704.766	31.72802	0
mod_dur	-3433.078	-3399.493	1723.539	37.54595	0

Table 7: Model summary for duration.

The **Akaike Information Criterion (AIC)** and **Bayesian Information Criterion (BIC)** were extremely large negative numbers for each nested model, indicating that we have a well-fit model. The **log-likelihood (logLik)** values were also very large, further indicating that the model is well-fit to the data. The values for **Chisq** were high; the values for the additive model and the whole duration model were in the 30s, but the Chisq value for the dur_island model was much larger at 86.22701. These values quantify the difference in fit between the models being compared. The greatest difference was for the island parameter.

Lastly, the **Pr(Chisq)** values for each model (excluding the null model) were 0. As this is less than 0.05, this suggests that the most complex model was a better fit than the nested models, which means the extra effects taken into account in the larger model were necessary.

5.2 Intensity

Data were analyzed using a linear mixed effects model (as implemented by the lme4 (version 1.1-37) package in R 4.5.0 (2025-04-11 ucrt)). This model included *intensity* as the dependent variable, and *island*, *speaker* and *word* as fixed factors.

The model summary fixed effects for the individual intensity model is shown in Table 8. The **standard error (SE)** for each island were large, around 11-15, indicating that the mean of the sample is not an accurate reflection of the actual populations mean. The **t-values** for each island were all lower than 2, indicating that they are not statistically significant between the islands.

	betas	SE	t-val	df
Atiu	57.919	10.864	5.331	6.184
Mauke	2.534	12.137	0.209	6.165
Penrhyn	-7.853	15.347	-0.512	6.156
Rarotonga	2.095	11.947	0.175	6.261

Table 8: Intensity fixed effects

The **degrees of freedom (df)** for all islands were around 6. For Mauke, **beta** was 2.534, which indicates that being a speaker from Mauke, intensity is expected to be higher. For Penrhyn, beta was -7.853, which indicates that being a speaker from Penrhyn, intensity is expected to be much lower. For Rarotonga, beta was 2.095, which indicates a speaker from Rarotonga is expected to have higher intensity, similar to Mauke. Atiu had the largest positive value for beta at 58, indicating the highest linear relationship between being a speaker from Atiu and having higher intensity.

As for the random effects, which are reported in Table 9, **standard deviation** for word and speaker were higher for the intensity model. For word, standard deviation was 4.377, so the data for intensity by word were less clustered around the mean. The standard deviation was even higher for speaker, at 10.834, so the speaker data was even more spread out. **Variance** values were consistent with this, as the variance for word was high at 19.16, and the variance for speaker was much larger at 117.38.

group	name	Std. Dev.	Variance
word	(intercept)	4.377	19.16
speaker	(intercept)	10.834	117.38
Residual		5.851	34.24

Table 9: Intensity random effects

Nested model comparison.

The model summary for the nested model comparison is in Table 10. For this, I utilized the ‘anova()’ function in R and omitted the columns I don’t interpret.

The intensity data were analyzed using a linear mixed effects model. Intensity was the criterion with *island*, *speaker* and *word* as predictors. Main effects and the interactions between *island*, *speaker* and *word* were assessed using nested model comparisons. Experiment-wise, alpha was set at 0.05.

	AIC	BIC	logLik	Chisq	Pr(>Chisq)
intensity_null	6613.473	6623.065	-3304.737	NA	NA
intensity_island	6500.268	6524.246	-3245.134	119.20555	0
intensity_add	5977.433	6006.207	-2982.717	524.83480	0
mod_intense	5902.351	5935.921	-2944.176	77.08165	0

Table 10: Model summary for intensity.

The **AIC** and **BIC** were extremely large numbers for each nested model, indicating that we do *not* have a well-fit model. The **log-likelihood** (logLik) values were very small, further indicating that the model is not well-fit to the data. The values for **Chisq** were extremely high; the values for the island model were 119.20555, the value for the additive model was 524.83480, and the value for the whole intensity model was 77.08165. These values quantify the difference in fit between the models being compared. The greatest difference was for the additive model, which included *intensity*, *island* and *speaker* parameters, excluding the *word* parameter.

Lastly, the **Pr(Chisq)** values for each model (excluding the null model) were 0. As this is less than 0.05, this suggests that the most complex model was a better fit than the nested models, which means the extra effects taken into account in the larger model were necessary.

5.3 F0

Data were analyzed using a linear mixed effects model (as implemented by the lme4 (version 1.1-37) package in R 4.5.0 (2025-04-11 ucrt)). This model included *intensity* as the dependent variable, and *island*, *speaker* and *word* as fixed factors.

The fixed effects for the individual F0 model are reported in Table 11. The **standard error (SE)** for each island were very large, at least 24, indicating that the mean of the sample is not an accurate reflection of the actual populations mean. The **t-values** for each island were all lower than 2, indicating that they are not statistically significant between the islands. The **degrees of freedom (df)** for all islands were around 3.4.

For each island, the **betas** are very spread out. For Mauke, beta was about 24, Penrhyn's beta was -3, and Rarotonga's beta was nearly 50. This indicates that a speaker from Mauke or Rarotonga is expected to have an increased f0 value (the increase is greater for Rarotonga). But, for Penrhyn, the f0 is expected to be lower for a speaker from Penrhyn. Again, Atiu has the highest positive number for beta, at 141, indicating a strong linear relationship between being from Atiu and having a higher f0 value.

	betas	SE	t-val	df
Atiu	140.935	24.584	5.733	3.466
Mauke	24.229	27.450	0.883	3.448
Penrhyn	-3.145	34.660	-0.091	3.423
Rarotonga	49.950	27.269	1.832	3.564

Table 11: F0 fixed effects

As for the random effects, reported in Table 12, the **standard deviations** for word and speaker were somewhat large, with standard deviation for word being 11.41 and standard deviation for speaker being 24.34. This indicates that the data for both was spread out, but more so for speaker. **Variance** for both word and speaker were also large, with variance for word being 130.1, and variance for speaker being 592.5. So, variability in this model was quite high.

group	name	Std. Dev.	Variance
word	(intercept)	11.41	130.1
speaker	(intercept)	24.34	592.5
Residual		28.67	822.2

Table 12: F0 random effects

Model fit and assumptions.

Since we have f0 values at three different intervals of the same morpheme, if we were to put all of those formant values in a new column and the percentage in another new column, then it would fabricate 3 times as many tokens that were not elicited from the speakers.

In order to limit this *autocorrelation* issue, I chose to only keep f0 values at the 50% interval of the phoneme. I chose the 50% interval because the middle of the duration seemed like the smartest choice to represent the average f0 value across the phoneme's duration.

Nested model comparison.

The model summary for the nested model comparison is in Table 13. For this, I utilized the 'anova()' function in R and omitted the columns I don't interpret.

The f0 data were analyzed using a linear mixed effects model. F0 (Hz) was the criterion with *island*, *speaker* and *word* as predictors. Main effects and the interactions between *island*, *speaker* and *word* were assessed using nested model comparisons. Experiment-wise, alpha was set at 0.05.

	AIC	BIC	logLik	Chisq	Pr(>Chisq)
f0_null	8419.866	8429.335	-4207.933	NA	NA
f0_island	8243.201	8266.874	-4116.601	182.66463	0.00e+00
f0_add	8153.415	8181.822	-4070.707	91.78646	0.00e+00
mod_f0	8139.651	8172.793	-4062.826	15.76368	7.18e-05

Table 13: Model summary for f0

The **AIC** and **BIC** were extremely large numbers for each nested model, indicating that we do *not* have a well-fit model. The **log-likelihood** (logLik) values were very small (negative) numbers, further indicating that the model is not well-fit to the data.

The values for **Chisq** were somewhat high; the values for the island model were 182.66, the value for the additive model was 91.79, and the value for the whole intensity model was 15.76. These values quantify the difference in fit between the models being compared. The greatest difference was for the island model, which included *f0* and *island* parameters, excluding the *speaker* and *word* parameters.

Lastly, the **Pr(Chisq)** values for each model (excluding the null model) were extremely small. As they are all less than 0.05, this suggests that the most complex model was a better fit than the nested models, which means the extra effects taken into account in the larger model were necessary.

5.4 F2

Data were analyzed using a linear mixed effects model (as implemented by the lme4 (version 1.1-37) package in R 4.5.0 (2025-04-11 ucrt)). This model included *f2* (Hz) as the dependent variable, and *island*, *speaker* and *word* as fixed factors.

The fixed effects for the individual f2 model are reported in Table 14. The **standard error (SE)** for each island were very large, at least 83, indicating that the mean of the sample is not an accurate reflection of the actual populations mean. The **t-values** for each island were all higher than 2, indicating that they *are* statistically significant between the islands.

	betas	SE	t-val	df
Atiu	1142.411	83.412	13.696	5.643
Mauke	297.881	90.354	3.297	4.946
Penrhyn	384.930	111.231	3.461	4.466
Rarotonga	389.099	94.770	4.106	6.027

Table 14: F2 fixed effects

The **degrees of freedom (df)** for Mauke and Penrhyn were under 5 and the df for Rarotonga was a bit above 6. For each island, the **betas** are very large. For Atiu, beta was 1142.41, Mauke's beta was 297, Penrhyn was 384, and Rarotonga was 389. This indicates all islands are associated with a large increase in F2, with the highest being Atiu.

As for random effects, reported in Table 15, **standard deviations** for word and speaker were quite high again. For word, the standard deviation was high at 193.80, and for speaker it was at 68.01. This indicates that for both word and speaker, the data was quite spread out, but the data for word was much more spread out around the mean.

group	name	Std. Dev.	Variance
word	(intercept)	193.80	37558
speaker	(intercept)	68.01	4625
Residual		394.20	155390

Table 15: F2 random effects

Variability was also very large for word and speaker in the F2 model, with variance for word being

37558 and variance for speaker being 4625.

Nested model comparison.

The model summary for the nested model comparison is in Table 16. For this, I utilized the ‘anova()’ function in R and omitted the columns I don’t interpret.

The f2 data were analyzed using a linear mixed effects model. F2 was the criterion with *island*, *speaker* and *word* as predictors. Main effects and the interactions between *island*, *speaker* and *word* were assessed using nested model comparisons. Experiment-wise, alpha was set at 0.05.

The **AIC** and **BIC** were extremely large numbers for each nested model, indicating that we do *not* have a well-fit model. The **log-likelihood** (logLik) values were very small (negative) numbers, further indicating that the model is not well-fit to the data.

	AIC	BIC	logLik	Chisq	Pr(>Chisq)
f2_null	13451.07	13460.66	-6723.536	NA	NA
f2_island	13394.48	13418.46	-6692.240	62.592574	0.0000000
f2_add	13394.49	13423.27	-6691.246	1.987579	0.1585942
mod_f2	13330.46	13364.03	-6658.229	66.034384	0.0000000

Table 16: Model summary for f2.

The values for **Chisq** were somewhat high, except for the additive model. The Chisq value for f2_island was high at 62.59, and the value for the full mod_f2 was 66.03. The additive model’s Chisq value was low at 1.99. These values quantify the difference in fit between the models being compared. The greatest difference was for the island model and the full model. The island model included *f2*, *island* parameters and the full model included *f2*, *island*, *speaker*, and *word*. Since the additive model was so low, the *speaker* parameter didn’t indicate a huge difference in the model fit.

Lastly, the **Pr(Chisq)** values for each model (excluding the null model) were extremely small, but the additive model was slightly larger at 0.16. Since the additive model was greater than 0.05, this indicates that the less complex models were a better fit than the more complex ones in the nested model comparison.

5.5 Interim summary

So, what do all of these models and comparisons tell us about the w–v alternation in CIM? First of all, the duration models were well-fit and accurate to the population and the nested model comparison showed that the island a speaker was from made the greatest difference in duration. Penrhyn had the largest mean duration and Rarotonga mean had the smallest duration.

Second, Atiu had the highest intensity and Rarotonga and Mauke had fairly high intensity values. The nested model was not well-fit, but the greatest difference in the nested model comparison was for the model with all parameters except *word*. So, intensity was more varied for *island* and *speaker* and *word* was not as relevant in this comparison.

Third, the F0 model was not well-fit and had high variability. The islands Atiu and Rarotonga were more associated with higher F0 values. In the nested model comparison, the biggest difference was for the model with only *island* and *F0*, though the model was not well-fit.

Finally, the F2 model was statistically significant and all islands had a large F2 value, with the highest being Atiu. The nested model comparison was not well-fit and the complex models were worse-fit. Speaker did not seem to make a big difference in model fit.

From these models and nested models, we can predict that Atiu and Rarotonga have more instances of voiced w–v phones (due to the F0 model), and Atiu should have the most w–v tokens with a high F2, so the phones should surface as [v] most often for speakers from Atiu. Duration also varied robustly between islands, and Penrhyn was associated with the highest durations and Rarotonga was associated with the smallest duration.

The intensity models did not seem to vary in a meaningful way.

5.6 Formant models

Recall that we had the following p-values in each model for each island.

Table 17: P-values from each model.

Island	Duration	Intensity	F0	F2
Atiu	0.00225	0.00161	0.00702	0.0000152
Mauke	0.90982	0.84133	0.43464	0.0219000
Penrhyn	0.32459	0.62670	0.93276	0.0216200
Rarotonga	0.25014	0.86632	0.14965	0.0062600

In the duration model, only Atiu did *not* have a statistically significant (> 0.05) p-value (0.00225), but the other three islands had statistically significant p-values. Penrhyn and Rarotonga had very similar p-values but Mauke had a significantly higher p-value. In the intensity model, again, all of the islands had statistically significant p-values (> 0.05) except Atiu. The highest p-values amongst the other three islands was for Make and Rarotonga. The model for F0 revealed that each island except for Atiu had a significant p-value. Penrhyn had the highest p-value in the F0 column.

Recall that an F0 value lower than **100 Hz** indicates that the phoneme is voiceless. We can tidy the F0 dataset once again to add a new column stating whether the w~v in that row surfaces as voiced or voiceless. The counts of voiced and voiceless tokens for each island are summarized in the following table.

Table 18: Voiced and voiceless phoneme counts.

island	Atiu	Mauke	Mauke	Penrhyn	Penrhyn	Rarotonga	Rarotonga
voiced	voiced	voiced	voiceless	voiced	voiceless	voiced	voiceless
freq	114	404	14	167	1	139	2

With the count information, we can manually calculate the percentages of voiceless w~v tokens surfacing in each group (island) of the data set. So, to answer the voicing question, only Atiu had zero voiceless tokens surfacing. Mauke had a voiceless surface form 3.10% of the time, which was the highest percentage of voiceless surface forms in the data set. Penrhyn and Rarotonga had similar rates of voiceless surface forms, with voiceless phonemes surfacing 0.59% of the time for Penrhyn, and 1.28% of the time for Rarotonga.

Finally, the F2 model revealed no statistically significant differences in F2 values between these four islands, as all of the p-values in the F2 column are less than 0.05.

The final research question for us to address is the w~v question. An F2 value **less than 850 Hz** is taken to indicate a [w] phoneme on the surface instead of a [v]. To answer this question in percentages, we can create an additional column that states whether F2 is greater or less than 850 Hz. If F2 is less than 850, the value in the new column is “w”, and if it is greater than 850, the value in the new column is “v”.

Once we have the counts of “w” and “v” surface forms for each island, we can calculate the percentages. The table with the frequency column is shown below.

Table 19: Frequency counts.

island	Atiu	Atiu	Mauke	Mauke	Penrhyn	Penrhyn	Rarotonga	Rarotonga
surface_form	v	w	v	w	v	w	v	w
freq	97	21	424	26	163	7	151	5

I manually calculated the percentages of [w]s surfacing in each group (island) based on the frequencies in the table above. The percent of [w]s in Atiu was 17.80%, 5.75% for Mauke, 4.12% for Penrhyn and 3.21% for Rarotonga.

6 Conclusion

Through doing this statistical analysis of w~v tokens across Atiu, Mauke, Penrhyn, and Rarotonga, we found that duration, intensity, and F0 by island are statistically significant for all islands except Atiu. Duration was the most significant for Mauke, and intensity and F0 were the most significant for Penrhyn.

The F2 model revealed no statistical significance between the F2 values and which island a speaker was from. However, to answer the voiced or voiceless question, we calculated percentages that show that all islands except Atiu had voiceless surface forms for w~v, with Mauke having the highest percentage of 3.10%.

In order to answer the final question, which islands have more [w] surface forms, we also calculated percentages. These percentages showed that all of these four islands had [w] phonemes realized, but the highest percentage of [w]s surfacing was in Atiu, where [v] was realized as [w] 17.80% of the time.

References

- [1] Coto-Solano, Nicholas & Wray. 2018. Development of Natural Language Processing Tools for Cook Islands Māori. In *Proceedings of Australasian Language Technology Association Workshop*, pages 26-33.
- [2] Nicholas, Sally Akevai (collector), 2012. Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects) . Collection SN1 at catalog.paradisec.org.au [Open Access]. <https://dx.doi.org/10.4225/72/56E9793466307>
- [3] Thieberger & Barwick. 2012. Keeping records of language diversity in melanesia, the pacific and regional archive for digital sources in endangered cultures (paradisec). *Melanesian languages on the edge of Asia: Challenges for the 21st Century*, pages 239-53.