Age and Gender Prediction in the DAMP Archive

Youkow Homma
Faculty Sponsor: Prof. Alan Gous
Mentor: Prof. Perry Cook

1 Introduction

Smule is a tech company based in San Francisco that builds mobile phone apps allowing its users to collaborate and socialize through musical performances. One of Smule's applications is a social karaoke app called "Sing! Karaoke" through which its users can sing along to the original backtrack of thousands of songs and collaborate with other users. When a person performs on Sing!, Smule maintains both metadata about the singer and performance as well as a vocal recording of the performance. Some of this data has been allocated for research purposes by the Smule MIR and Data Science groups as the Stanford Digital Archive of Mobile Performances (DAMP) [Smi13]. This database includes Sing! data from two songs: Amazing Grace in the style of Elvis Presley and Let It Go from the Disney movie, *Frozen*, where each entry in the dataset corresponds to a performance on the app.

Each data entry contains metadata about the singer's self-declared gender, self-declared age, country, language, headphone usage, and recording date, as well as an audio recording of the vocal track of the performance. The Amazing Grace dataset contains roughly 17,000 records while the Let It Go dataset is significantly larger with about 240,000 records. Of these records, the Amazing Grace dataset has about 7,500 records missing both gender and age metadata while the Let It Go dataset has about 130,000 entries missing both gender and age. Thus, in both instances, roughly half of the records are missing key information about the singer. Morever, since both gender and age are self-declared, even the populated entries are sometimes inaccurate.

In this project, we study the user and vocal data for these two songs with the aim of filling in the missing gender and age metadata as well as validating the existing data. We also prototype a music information retrieval (MIR) pipeline that could be implemented for extracting audio features from each audio recording. By getting a more complete view of its users, Smule can make better recommendations for songs to sing and also better connect its users to recordings made by others. Furthermore, the MIR pipeline results will allow Smule to understand the audio and vocal characteristics that are important for completing this missing data.

The remainder of this paper is structured as follows: In section 2, we look at the relationship between the metadata we want to predict (age and gender) and the other metadata. In section 3, we discuss the relationship between age and gender and the vocal features extracted from the audio. Both of these sections make frequent reference to the figures contained in section 8. In section 4, we look at the prediction results for age and gender and discuss what features are most important for prediction. In section 5, we discuss how the results can be applied at Smule, and in section 6, we discuss the work that was performed for this project from start to finish and also suggest what more can be done in the future.

2 Exploring the Metadata

First, we'll start by looking at the relationship among the response variables (gender and age) and the other metadata relating to geography and creation of the audio recording. More specifically, this additional metadata is:

- Country code - Country associated with the device
- Locale - the country and language associated with the singer's account
- Latitude/Longitude/City Id - Geographic location of the performance
- Creation Timestamp - Date of the performance

For each of the datasets, we perform some pre-processing steps to ensure that each dataset is sensible. These steps are as follows:

1. In each dataset, there are many performances by the same user leading to duplicate metadata. For instance, there is one user in the Let It Go dataset that has performed the song an astounding 85 times. Overall, there are 14,398 unique singers for Amazing Grace and 160,809 unique singers for Let It Go, meaning an average of about 1.2 and 1.5 performances per user, respectively. In the following analysis, we consider each user only once by using the metadata associated with the first performance for each user.
2. As mentioned before, the age is self-reported and has a chance of being inaccurate. Some of these inaccuracies are easy to spot. For instance, there are about 30 accounts in the Amazing Grace dataset with an age over 100. In order to minimize the effect of these accounts and narrow our analysis to mostly correct information, we only consider users with age at most 75.
3. We further augment and modify the metadata in the following ways:
   - Map the country code to a country name using the name used by the European Central Bank and consider this the country associated with the performance. An alternative is to use the geographic location of the performance via the city id; however this field is missing for about two-thirds of the data. We will discuss this problem further in section 6.
   - Map the country to a geographic region as identified by the World Bank Development Indicators.
   - Extract the language code from the locale and map it to a name using the ISO standard.
   - Truncate the creation timestamp at the date and ignore the time.
   - We calculate the user's age as the year of the performance minus the birth year

With these modifications, we will now analyze the metadata of the two songs separately. Throughout the discussions, we will make frequent reference to the figures at the end of this paper in section 8.

2.1 Amazing Grace

Of the 7764 accounts in the Amazing Grace dataset that specified their gender, 2357 are male and 5407 are female, giving us roughly one-thirds male and two-thirds female singers. The median age for both genders is 30, and we see below in figure 1 that the distribution of age is fairly similar among males and females.

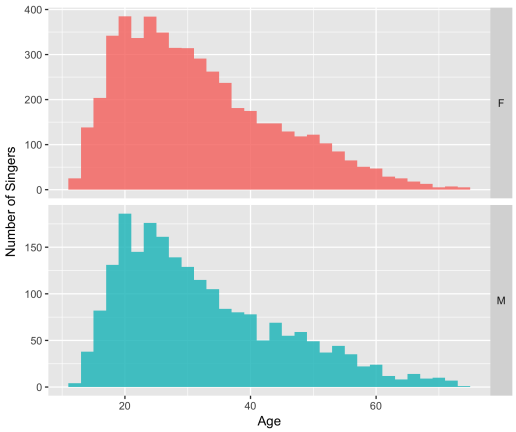

Figure 1: The distribution of age among male and female singers in Amazing Grace.

Overall, the Amazing Grace singers are relatively homogenous: 66% of the singers are in Northern America and 14% of the singers are in Northern Europe. Furthermore, 85% of the singers have their language set to English, while the second most popular language is French at 3% followed by German and Portugese at 2% each. Thus, the performances of this dataset is concentrated predominantly in the Western hemisphere.

In figure 5, we can see the proportion of males associated with each region and language. One interesting occurrence is the very high proportion of males in South America, as well as the relatively high proportion of males speaking Portuguese. Aside from these larger deviations, there is not much variability in the gender distributions among geographic splits.

Figure 6 demonstrates again that the populations are relatively similar across geographic distinctions, this time for age. For this figure, we handpicked a few distributions that had high significance when compared to

other distributions under the Wilcoxon test. South American singers tend to be slightly younger, as do French and Swedish speakers in France and Sweden, respectively, while English speakers in the Philippines tend to be older than singers in the US. Given that these are the extremely different age distributions, we can see there is also not much difference overall.

Figure 7 shows perhaps the most variability among the relationships of the metadata variables. Again, we handpick for the graphic some of the most different distributions. We can see that in Northern America and generally in the Western hemisphere, there is a gradual increase in performances from 2013 to 2016. On the otherhand, in other parts of the world like in South-Eastern Asia and among Indonesian speakers in Indonesia, almost all of the performances are concentrated after mid-2015. Finally in figure 8, we can see that there is a positive relationship between performance date and age. That is, the later a person has performed, the older that person is likely to be.

2.2 Let It Go

Of the 64,402 accounts in the Let It Go dataset that specified their gender, 13,067 are male and 51,335 are female, giving us roughly one-fifth male and four-fifths female singers. The median age for both genders is 25, and we see below in figure 2 that the distribution of age is again fairly similar among males and females. One key difference between Let It Go and Amazing Grace is that the singers for Let It Go are significantly younger - on average about five years. This makes age prediction easier as the range of ages is narrower for Let It Go, but also makes the data more unreliable as we will discuss more in section 4

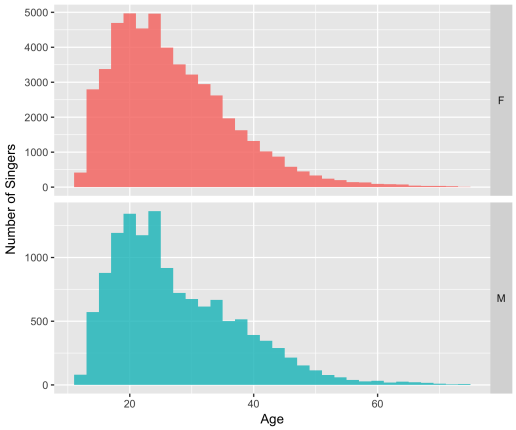

Figure 2: The distribution of age among male and female singers in Let It Go.

The Let It Go dataset is far more diverse than Amazing Grace due to both Disney's strong marketing for the movie *Frozen* as well as the almost 15 times as many records in the Let It Go dataset. In figure 9, we can see that there is a wide range for the proportion of males based on the geographic segmentation. For the regional breakdown, we see that Southern Asia has a significantly high proportion of male singers, as does South America which was also the case in Amazing Grace. Looking more closely at the countries in Asia in the middle chart, we see that countries we might expect to be similar based on geographic proximity end up on different ends of the graph. For instance Japan and South Korea have a significantly different proportion of male singers, as do the pairs Hong Kong/Taiwan and Singapore/Thailand. Finally, we see in the last graph that for the European-based languages, the Germanic languages (English, German, Dutch, Swedish, Norwegian) tend to have a smaller proportion of male singers than the Romance languages (Spanish, Portugese, French, Italian). Overall, there seems to be a great diversity in the proportion of male singers based on geographic and linguistic distinctions.

Similarly, the distribution of age also varies noticeably more than Amazing Grace based on the geographic region as shown in figure 10. There are far fewer children singing in South-Eastern Asia while there are far more children singing in Eastern and Western Europe when compared to Northern America and Northern Europe. We can see however in the right panel that language, not just geographic location plays a key role in the age distribution. For instance, we can see that the distribution of age for English speakers in the United States is very different from that of English speakers in Indonesia and Japan. However, these distributions in turn are different from the age distributions for Indonesian speakers in Indonesia and Japanese speakers in Japan. In fact, for the examples presented here, it seems the distribution for those speaking in their native tongue are younger than those where the locale's language is not the primary language spoken in the locale's country. We can also see this reflected in the French speakers from France.

Finally, we see again that the creation timestamp is the most diverse aspect of the metadata. Figure 11 shows that there is early adoption of Let It Go in Northern America with a peak in mid-2014, whereas in Central America and Asia, there is a large spike in performances in mid-late 2015. In the right panel, we can verify that this phenomenon is driven primarily regionally as English speakers in Indonesia and the Philippines also had a boom in performances in the latter part of 2015. In figure 12, we see that like in Amazing Grace there is a positive relationship between performance date and age.

3 Audio Features

For each performance, we have along with the metadata a recording of the vocal track of the performance. What's unique about this data set is that because the recordings are taken from a karaoke-style application, users that use headsets generally have minimal background noise, and the recordings only contain the voice of the singer. This makes analysis significantly cleaner than if we had to extract the vocal audio mixed with the background track using signal processing.

3.1 Audio Processing Pipeline

In order to analyze the audio, we use vocal analysis software implemented in MATLAB called VoiceSauce [YSY11] which was developed by the UCLA Phonetics Lab. We explored some other audio analysis software to prototype initial models, and we will discuss those directions further in section 6. For each recording of each song, we analyze the same 1.5 second long portion of the song and calculate a feature vector every 10ms. We then aggregate these feature vectors across time for each recording to get a mean and standard deviation for each feature for the recording. The features calculated by VoiceSauce are:

- 1st, 2nd and 4th Harmonics (H1, H2, H4)
- Harmonics nearest 2kHz and 5kHz (H2k, H5k)
- Spectral slopes (differences) between harmonics above
- Harmonic-to-Noise Ratios (HNR)
- Subharmonic-to-Harmonic Ratio (SHR)
- Cepstral Peak Prominence (CPP)
- Energy
- Fundamental Frequency (F0)
- Strength of Excitation

When picking which recordings to analyze, it was important to filter out songs that would yield a usable sample since each recording is expensive to analyze. For Amazing Grace, we were fortunate that Professor Perry Cook had already pre-analyzed some recordings and determined whether or not the recording had an abundance of background noise. Thus, for Amazing Grace, we only analyzed those songs which were deemed to be clean in this pre-analysis. For Let It Go, no such pre-analysis had been performed, so we simply filtered for songs that had been recorded with a headset. Then, for both songs, we also eliminated after analysis, all recordings with a low energy (mean energy less than 0.02) since almost every song listened to below this level had no perceivable vocal audio.

In Amazing Grace, we analyzed all recordings from seconds 24 to 25.5 during the first instance of "like me..." in the lyrics. In Let It Go, the 1.5 second clip is located from time 59 to 60.5 during the chorus' "let it go..." On my local machine, I was able to process each 1.5 second clip in about 30 seconds, while on the ICME cluster, it took about 40 seconds. In total, about 3,500 total recordings were processed for Amazing Grace and 6,200 recordings were processed for Let It Go.

4 Metadata Prediction

We now tackle the problem of predicting gender and age for the Amazing Grace and Let It Go datasets using the vocal features described above and the metadata from section 2. In order to study the effectiveness of adding an MIR pipeline, we perform predictions of age and gender under four settings:

1. Using only the original metadata, with no voice quality metrics
2. Using the original metadata along with the voice quality metrics except for the differences between raw harmonic values (spectral slopes)
3. Using the original metadata along with the voice quality metrics except for the raw harmonic values (H1, H2, H4, H2k and H5k)
4. Using the original metadata along with all features computed by VoiceSauce

By comparing setting 1 to any of the other three settings, we can see how much the voice quality metrics can improve prediction of the missing data. Settings 2 and 3 are used to draw better inferences about the importance of the features. Since the spectral slopes are linear combinations of the other predictors, putting both the raw harmonics and the spectral slopes in the model dilutes the importance metric when compared to other predictors.

On a similar note, some of the harmonic ratios have a very high correlation which can also dilute the importance metrics, so we remove some of these correlated pairs. In particular, in Amazing Grace, we remove HNR15 and HNR35 mean and standard deviation due to their high correlation with the HNR25 mean and standard deviation, repsectively. In Let It Go, we remove the HNR05, HNR25 and HNR35 means due to their high correlation with the HNR15 mean, and the HNR15 and HNR35 standard deviations because they are highly correlated with HNR25_sd. Finally, we also remove the Energy standard deviation in Let It Go due to its high corelation with the mean Energy.

For each of the models, we perform a ten-fold cross validation over the following set of model parameters:

- Penalized GLM:
  - $\alpha \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 0.3, 0.5, 0.7, 0.9, 1-10^{-1}, 1-10^{-2}, 1-10^{-3}, 1-10^{-4}, 1\}$
  - $\lambda \in \{1.8^{-1}, 1.8^{0}, \ldots, 1.8^{10}\}$
- Random Forest:
  - Number of trees = 120
  - Variables considered per split $\in \{\frac{\#predictors}{4}, \frac{\#predictors}{3}, \frac{\#predictors}{2}\}$ for age and $\{\sqrt{\#predictors}/1.5, \sqrt{\#predictors}, 1.5\sqrt{\#predictors}\}$ for gender
- GBM:
  - Shrinkage = 0.05
  - Number of trees $\in \{10, 25, 50, 90, 170, 300, 450\}$
  - Interaction depth $\in \{1, 2, 4\}$
  - Minimum observations per node $\in \{2, 5, 10\}$ for age and $\{1, 3, 5\}$ for gender

We select the final parameters by using the one standard error rule - that is, we choose the least complex model which has a cross-validation error that is still within one standard error of the error acheived by the best parameter setting. We choose this methodology to avoid overfitting to the noise in the cross-validation procedure.

4.1 Gender Classification

To set up our gender classification, we first need to define our training and test sets. As we noted in section 2, about two-thirds of the population is female in Amazing Grace and four-fifths is female in Let It Go. In order to account for this imbalance, we construct our training and test sets to be balanced with half male and half female recordings. Furthermore, we do not filter for repeat recordings - that is, there is some correlation between samples in the training and test set because the same singer may have multiple recordings. Although this leaks some of the testing information into the test set, we want to use the available analyzed data as much as possible due to its already modest size. This leaves us with a training set of 1650 recordings and test set of 180 recordings in Amazing Grace, and a training and test set of 2200 and 240 recordings, respectively, for Let It Go.

For gender classification, one easy method for classifying gender is to look at the mean fundamental frequency. In figure 3, we can see that in both songs, the distribution in fundamental frequency is bimodal with males singing at the lower frequency and females singing at the higher frequency. The split in Amazing Grace is a little neater than in Let It Go. For both songs, we find the optimal frequency split point in our training set, which is 278.5 for Amazing Grace and 269.1 for Let It Go. Using this split point on our test set then yielded a test error of 7.7% in Amazing Grace and 26.1% in Let It Go. We will consider this our baseline model and aim to beat these error rates when incorporating the other voice quality metrics to show that there is more information beyond pitch that can help with gender prediction.

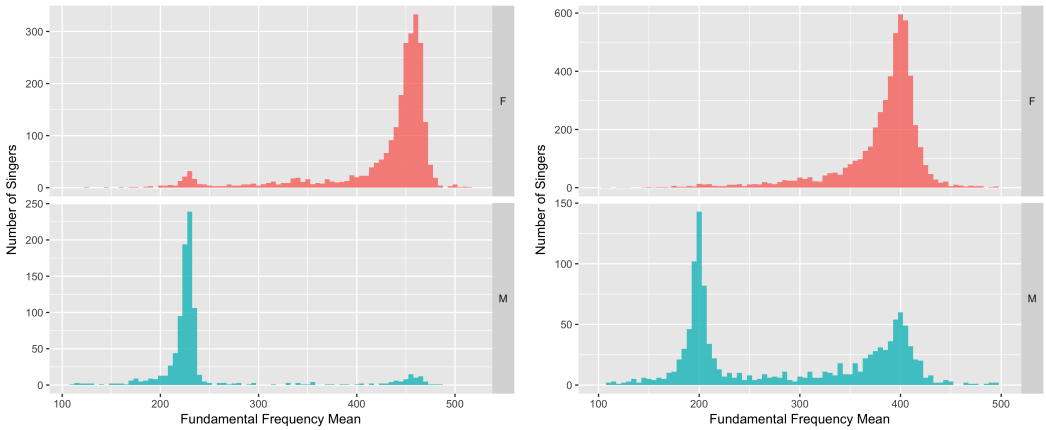

Figure 3: Fundamental frequency in Amazing Grace (left) and Let It Go (right) broken down by gender for all analyzed recordings.

Using the models and methodology discussed at the beginning of the section, we acheive the test classification results found in tables 1 and 2. We can see that in both cases, the models with voice quality data far outperform the models with only metadata. We can also see that the random forest outperforms the other models in both songs in all settings with vocal data. Furthermore, the 7.1% test error in Amazing Grace and the 21.0% test error in Let It Go when using the harmonic differences with a random forest both beat the baseline model of splitting along the mean fundamental frequency. Finally, we can see that the performance does not degrade, and in fact, is slightly better for random forests when we remove the raw harmonics. These comparisons suggest that these additional predictors are not necessary in the model.

Model	Metadata Only	Metadata + Raw Harmonics	Metadata + Harmonic Diffs	All Data
Penalized GLM	40.1%	8.8%	8.2%	9.9%
Random Forest	37.9%	7.1%	7.1%	7.7%
GBM	49.4%	7.7%	7.7%	7.7%

Table 1: Amazing Grace Gender Classification Error

Model	Metadata Only	Metadata + Raw Harmonics	Metadata + Harmonic Diffs	All Data
Penalized GLM	42.4%	23.9%	22.3%	22.3%
Random Forest	45.3%	21.4%	21.0%	22.7%
GBM	51.7%	24.4%	26.5%	22.7%

Table 2: Let It Go Gender Classification Error

Now, we take a look at the feature importances for the models. The random forest was superior to both the penalized GLM and the GBM, so we will consider the feature importances from the random forest models. The feature importances are computed using permutation importance, where the importance metric reported is the increase in the out-of-bag MSE when the feature's values are permuted.

In table 3, we can see that the top five most important predictors for Amazing Grace are the same for each of the three random forest models. We can take a look at the distributions of each of these predictors for each gender in figure 13. By visually inspecting the graphs, we can see why these variables are important for gender prediction. The standard deviation in CPP is bimodal with the two peaks corresponding to the different genders, much like for the mean fundamental frequency. The 05 harmonic-to-noise ratio mean is shifted slightly closer to 0 for males. Finally in both the strength of excitation and fundamental frequency standard deviations, males have much narrower distributions that are clustered closer to 0.

In table 4, we can see that the important predictors are a little more diverse for Let It Go; however, there is still significant overlap among the models. We again plot the distributions of some of the common predictors in figure 15. Encouragingly, we can see that aside from the mean fundamental frequency, there are three other predictors, CPP standard deviation, HNR05 mean and fundamental frequency standard deviation that are important in both songs. The most promising predictor in Amazing Grace was CPP standard deviation, and we can see in figure 15 that while the female distribution is similar between the two songs, the male distribution is not. On the

other hand, we can see that the HNR05 mean and funamental frequency standard deviation are similar between songs for both males and females. This suggests that these differences may persist and generalize to other songs as well. In section 5, we will talk about how these findings relate to other works that have used voice quality to classify gender.

Feature	Importance
F0 Mean	21.7
CPP Std. Dev.	12.8
F0 Std. Dev.	12.2
HNR05 Mean	11.2
SoE Std. Dev.	10.2

Feature	Importance
F0 Mean	29.0
CPP Std. Dev.	14.3
HNR05 Mean	12.2
SoE Std. Dev.	11.5
F0 Std. Dev.	11.1

Feature	Importance
F0 Mean	25.0
CPP Std. Dev.	13.7
HNR05 Mean	11.1
F0 Std. Dev.	10.9
SoE Std. Dev.	10.3

Table 3: Feature importances for the raw harmonics (left), harmonic differences (center), and all predictors (right) random forest models for Amazing Grace gender classifcation.

Feature	Importance
F0 Mean	19.2
SHR Std. Dev.	13.0
F0 Std. Dev.	11.0
HNR05 Mean	9.8
CPP Std. Dev.	9.5

Feature	Importance
F0 Mean	19.7
SHR Std. Dev.	13.0
F0 Std. Dev.	11.2
SHR Mean	9.6
HNR05 Mean	9.4

Feature	Importance
F0 Mean	21.5
SHR Std. Dev.	13.0
F0 Std. Dev.	9.5
H4-2K Mean	9.4
SHR Mean	9.4

Table 4: Feature importances for the raw harmonics (left), harmonic differences (center), and all predictors (right) random forest models for Let It Go gender classifcation.

Overall, we can see that we have a much more difficult time classifying gender in Let It Go as our best model still has a test error rate of 20%. After listening to the incorrect classifications, we found that about 40% of the misclassifications had metadata associated with someone between the ages of 35-60, and that 90% of those recordings had the voice of a child. This suggests that many children are singing using the metadata of their parent, making the training data extremely noisy. This noisiness is supported by the fact that our cross-validations tended to favor extremely low-complexity models. For instance, the cross-validation procedure for the GLM model chose $\alpha = 0$ (i.e. the sparsest model) for every training instance. This noisiness also explains why the GBM performs poorly while the random forest is much more robust. The GBM generally reduces the bias of the model by making the model more complex at each boosting iteration; however, with so much noise, the bias reduction seems to overfit too easily to the noise. On the other hand, the random forest reduces the variance of the model without overfitting to the noise.

4.2 Age Prediction

To set up our age predictions, we again need to define our training and test sets. In this instance, we don't have to account for an imbalanced data set and use all of the analyzed data available to us. In this instance, we aim to reduce the squared error loss between the predicted age and the age reported in the metadata. Again, we do not filter for repeat recordings, so there is some correlation between samples in the training and test set. After filtering for mean energy above 0.02, we end up with a training set of 3,000 recordings and test set of 330 recordings in Amazing Grace, and a training and test set of 5,800 and 310 recordings, respectively, for Let It Go.

For age prediction, we use a null model in each instance that simply predicts the mean age of the training set. In Amazing Grace, the mean age is 32.5, yielding a test MSE of 175.9, and in Let It Go, the mean age is 26.9, resulting in a test MSE of 89.9. This demonstrates that the age is already much easier to predict in Let It Go by virtue of the ages being highly clustered around a younger age. Now, in tables 5 and 6, we can see the test errors for our regression models. We again see that the random forest beats the other two models in both songs. This result is not surprising given the evidence of mislabeled data described in the previous subsection.

Model	Metadata Only	Metadata + Raw Harmonics	Metadata + Harmonic Diffs	All Data
Penalized GLM	175.3	162.7	163.0	161.2
Random Forest	215.0	159.7	162.7	158.5
GBM	178.4	174.6	175.6	174.1

Table 5: Amazing Grace Age Mean Squared Error

Model	Metadata Only	Metadata + Raw Harmonics	Metadata + Harmonic Diffs	All Data
Penalized GLM	88.0	86.0	85.9	85.7
Random Forest	110.0	84.8	83.4	84.7
GBM	90.7	90.0	90.6	90.0

Table 6: Let It Go Age Mean Squared Error

Looking at the feature importances of the random forest models in table 7 suggests that age of Amazing Grace singers is well-predicted by the H1-H2 mean, CPP standard deviation, H2K mean and H4-2K standard deviation. We can see the relationship of age with these predictors in figure 14. By examining the p-values, we can see that in general, the relationships between the predictors and age is stronger for females. Furthermore, the relationship for H2K mean seems rather weak, suggesting there may be other interaction terms that are important to using this variable.

It is difficult to draw inferences from the Let It Go age models, first and foremost because the models are not very strong - even the best models only slightly improve on the baseline model of predicting the mean. Moreover, there is not much overlap between the important features of the different models. Nevertheless, we note that the fundamental frequency standard deviation shows up in all three models while the H5K mean shows up in the raw harmonics and all predictors models. The relationship of age with these predictors is depicted in figure 16.

Feature	Importance
CPP Std. Dev.	21.7
H2 Mean	17.7
Energy Std. Dev.	15.8
H2 Std. Dev.	14.8
H2K Mean	14.4

Feature	Importance
H1-H2 Mean	11.7
CPP Std. Dev.	11.0
SoE Mean	8.5
H4-2K Std. Dev.	8.2
CPP Mean	8.2

Feature	Importance
H1-H2 Mean	13.0
CPP Std. Dev.	10.8
H2K Mean	8.7
H4-2K Std. Dev.	7.7
Energy Mean	7.4

Table 7: Feature importances for the raw harmonics (left), harmonic differences (center), and all predictors (right) random forest models for Amazing Grace age prediction.

Feature	Importance
H1 Mean	14.6
H5K Mean	11.3
H4 Std. Dev.	11.2
H1 Std. Dev.	10.9
F0 Std. Dev.	10.9

Feature	Importance
F0 Std. Dev.	10.9
CPP Std. Dev.	10.7
Perf. Date	10.1
CPP Mean	9.8
SHR Mean	9.6

Feature	Importance
F0 Std. Dev.	11.3
Language	11.2
H2K-H5K Mean	10.4
H2K Mean	9.8
H5K Mean	9.5

Table 8: Feature importances for the raw harmonics (left), harmonic differences (center), and all predictors (right) random forest models for Let It Go age prediction.

Unfortunately unlike in the gender classification case, there is not much overlap among the important features between the two songs. We can see that perhaps the standard deviation of Cepstral Peak Prominence may be an important predictor for age as there is a similar relationship to age in both songs. Furthermore, this feature stood out in our gender prediction models as well, so further study on this feature is merited.

One further observation is that in Let It Go, we have the appearance of metadata predictors, namely performance date and language. We saw in our metadata exploration that there seemed to be a high diversity in age distribution among geographic partitions (figure 10), and that there was a strong positive relationship between age and performance date (figure 12). These relationships seem important enough to stand out in the case of age prediction in Let It Go.

5 Discussion

From the results in the previous section, it is clear that gender and age prediction can be greatly enhanced with the addition of voice quality metrics. Moreover in doing so, we found that fundamental frequency was the most important factor in classifying gender; however, there were some othe common important predictors between the two songs and the two prediction problems, namely the CPP standard deviation, HNR05 mean and fundamental frequency standard deviation.

There is some literature to support our findings related to the important features in both age and gender prediction. Manjunatha et. al. [Man17] found that when young adults were asked to phonate vowels at their normal pitch, there was a main effect of age on CPP, but not of gender. Our results from Amazing Grace support that there is some relationship between age and gender even in singing voices, though the relationship is in the standard deviation of CPP. Another study by Chen et. al., [GCA10] used VoiceSauce to classify the gender of children using speech data and found that CPP, HNR and the H2-H4 spectral slopes improved gender classification accuracy when added to models using just the fundamental formant frequencies. In summary, our results for important predictors align with some of the existing literature for similar problems in the speech domain.

There are a few key points that need to be emphasized before these results are implemented in the wild. First, the gender classification study was performed on a balanced data set. As mentioned previously, the gender populations for the songs we studied are imbalanced towards females, so one should next expect the same classification accuracy. Furthermore, when we constructed the training and test sets, we removed any audio sample with a mean energy lower than 0.02. This was to ensure that the audio clip contained enough sound to be plausibly someone singing. In essence, we filtered for the cleaner recordings, so again, we would expect these models to perform worse on audio sampled from all available audio samples.

The merits of computing vocal features are clear; however, the implementation methods are less so. The basics steps we used here can be replicated at Smule as follows:

1. Identify songs that many people sang with a relatively balanced gender population
2. Identify portions of the song with a long, held out note (these are somewhat tricky to identify because if the note is too long, poorer singers will not be able to hold the note and the result will be silence)
3. Perform audio analysis on the recordings of the song and extract important features including fundamental frequency, harmonics, and CPP statistics
4. Train a random forest model on the desired metadata by finding the parameters using one standard error cross validation
5. Aggregate the metadata predictions across different songs to get a more accurate prediction for each user

When picking popular songs in step 1, the problem of mislabeled data that we encountered with Let It Go is almost certain to ocurr. When looking at the gender classification problem, we saw a hint of the extent to which this problem is prevalent. A more clear picture of this phenomen can be see when plotting the mean fundamental frequency against the age in the two songs, as shown in figure 4 below. We can see that in Amazing Grace, there is a strong negative relationship between age and frequency for both genders, as is to be expected since young children will sing at a higher pitch. On the other hand, in Let It Go, we see that for self-declared males, there is a positive relationship between age and pitch. In fact, a similar pattern appears to be true for CPP standard deviation when comparing figures 14 and 16.

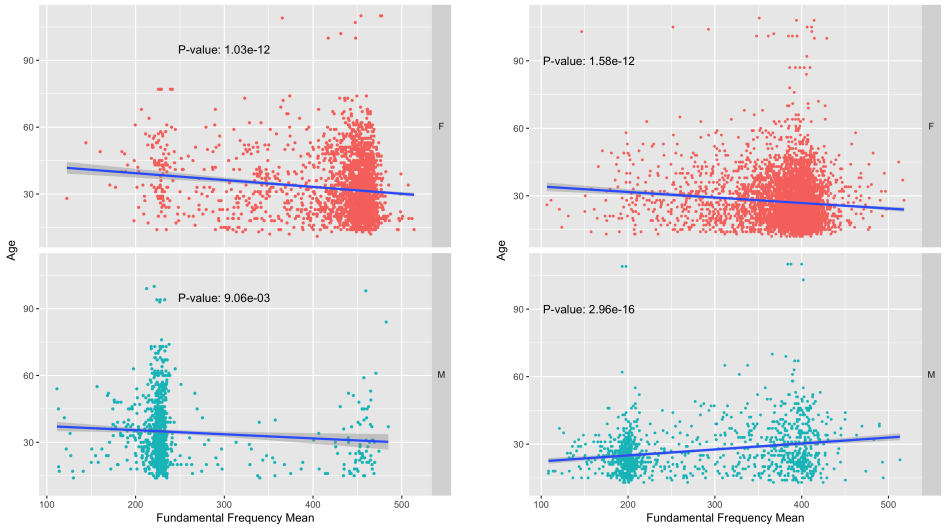

Figure 4: Mean fundamental frequency in Amazing Grace (left) and Let It Go (right) plotted against age and broken down by gender for all analyzed recordings.

In order to correct for this, there could be a separate model that flags accounts as potentially incorrect metadata using vocal features. As we saw with Let It Go, one way to do this is to examine songs with incorrectly classified gender for songs with many children singing.

Another consideration is the time it takes to analyze songs for step 3. In a production environment with millions of users, minimizing analysis time will be crucial to implementing an MIR pipeline. In particular, spending 30 seconds per user is simply not scalable. We experimented with shrink the amount of each song that was analyzed to 0.5 seconds, rather than 1.5 seconds in Let It Go; however, we found that the performance degraded significantly. We believe the problem stems from the fact that we focus the analysis on portions of the song with an extended pitch, so as we narrow the analysis window, there is a smaller chance that we actually capture the person singing because the singer either mistimes the note or has run out of breath.

Ultimately, the findings here point to how voice quality features can significantly add to the predictive power of models trained to identify user data, and we can identify specific features that help achieve this; however, there is still much care and thought required to implement such a training model into a production pipeline.

6 Work Performed and Future Work

The work performed for this project can be broken down into four parts. During the first phase, I explored the dataset to see which of the items listed in the project proposal and the introductory paper on DAMP [CS18] to study. While there is a plethora of geographic data that could have definitely led to some interesting projects, I decided to focus on the missing metadata problem because the goal was straightforward and I could still study some of the geographic data in the context of prediction. I then started some basic background reading to understand what kinds of vocal features were commonly looked at when predicting age and gender. During this search, I came across a blog post by Becker [Bec16] about extracting summary statistics about the vocal characteristics using the R package warbleR [ASSV17]. I used this as a launching point to use warbleR to get a preliminary analysis and baseline for the prediction problems. The warbleR package was limited in that it consisted almost solely of summary statistics of the distribution of the fundamental frequency. Moreover, when compared to the results from VoiceSauce, the warbleR metric didn't seem to be very accurate. Nevertheless, this work helped to establish an outline for the remainder of the project.

In the second phase, I explored how to analyze the audio files for voice quality metrics. I considered using Praat, Marsyas and VoiceSauce, but ultimately settled on VoiceSauce because it was implemented in MATLAB with an easy GUI, rather than the other two, which require their own scripting languages. After settling on VoiceSauce, I began analyzing the songs which took a lot more time and effort than I had anticipated because the analysis takes about 30 seconds per 1.5 second clip on my laptop and 40 seconds on the ICME cluster, meaning that with about 10,000 clips analyzed in total, the compute time was about 90 hours. Furthermore, VoiceSauce crashes and stops analyzing if it encounters a corrupted file or a file that has an empty channel. It

took a few overnight crashes to identify this as the source of the problem and to find a way to filter out these files before analysis. Finally, although the ICME cluster is a remote cluster, VoiceSauce uses a GUI interfaces, so the analysis will continue to run only as long as the connection to the cluster is kept open. During this phase, I learned to write some shell scripts as well as how to use the command line audio tool, SoX Sound eXchange.

In the third phase of the project, I built out the machine learning models using the metadata and analyzed recordings. To do so, I first cleaned the data by converting all of the codes to names using the ISO standard and the R package countrycode [ABY18] and then performed the data transformations detailed in section 2. I then trained the models on this data by leveraging the packages rpart [TTR18], glmnet [JF18], randomForest [LW18] and gbm [Rid17] as well as caret [Kuh18] to cross-validate the hyper-parameters. The data cleansing allowed me to dig into the finer details of all the metadata and I began to notice some patterns as a result. The model-building was relatively straight-forward, although some care was needed when setting up the inputs to the model for deriving sensible inferences.

In the fourth and final phase of the project, I analyzed the results of the models, and constructed a cohesive narrative about the data using visualizations in R. Between all four phases, I ended up writing about 2000 lines of R code with the visualizations being a bulk of this. I also wrote this paper and created the poster presentation for the ICME Xpo. This paper was a great experience as it gave me a chance to organize everything I have done this quarter and tell a story with the data and results.

There are many directions that one could take with this project in the future. To build directly off of this work, it would be interesting to see whether extracting more VoiceSauce features could help prediction further. In particular, VoiceSauce actually has the capability to incorporate Praat and also to calculate formants. I did not incorporate these features due to time constraints, but it might be nice to see if they add anything to the models.

I think that the geographic exploration could still be a very interesting study of the dataset. In particular, as I mentioned earlier in the paper, almost half of the data is missing a city_id; however, most of the data has a longitude and latitude. Moreover, there are three geographic attributes in the metadata (longitude/latitude/city_id, country, and locale) that sometimes give very different pieces of information. For instance, I have found examples in the data where the city_id for the performance corresponds to San Francisco, the country is Taiwan and the locale is Indonesian language/Indonesia. One key step that would need to be taken here is to map the longitudes/latitudes to a country. I looked briefly at doing this through Google's Maps API; however, it seems that there is a call limit that is far too low for the about 250,000 performances in Let It Go. One idea here is to filter out all of the examples that lie in the US, but even then, the number may still be too large.

On a similar note, I wasn't able to study the relationship between the vocal metrics and the geography as much. In order to do so, it would be necessary to analyze more data from more geographic regions as the current set of analyzed recordings is still heavily concentrated in the US. I think that such a study could yield some interesting insights about regional singing characteristics.

7 Acknowledgements

Thanks to Prof. Perry Cook and Smule for providing data and weekly guidance throughout the quarter, Prof. Alan Gous for helpful suggestions and feedback, and Kari Hanson and ICME for providing computing resources and guidance.

8 Figures

8.1 Metadata Figures for Amazing Grace

8.1.1 Gender Proportions by Geographic Attributes

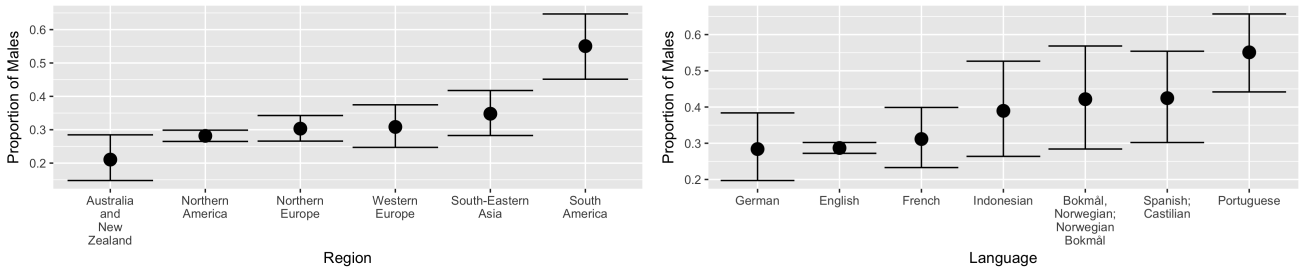

Figure 5: The proportion of males for each region/language with 99% confidence intervals, where the number of people for each region/language is at least 100 and 80, respectively.

8.1.2 Age Distributions by Geographic Attributes

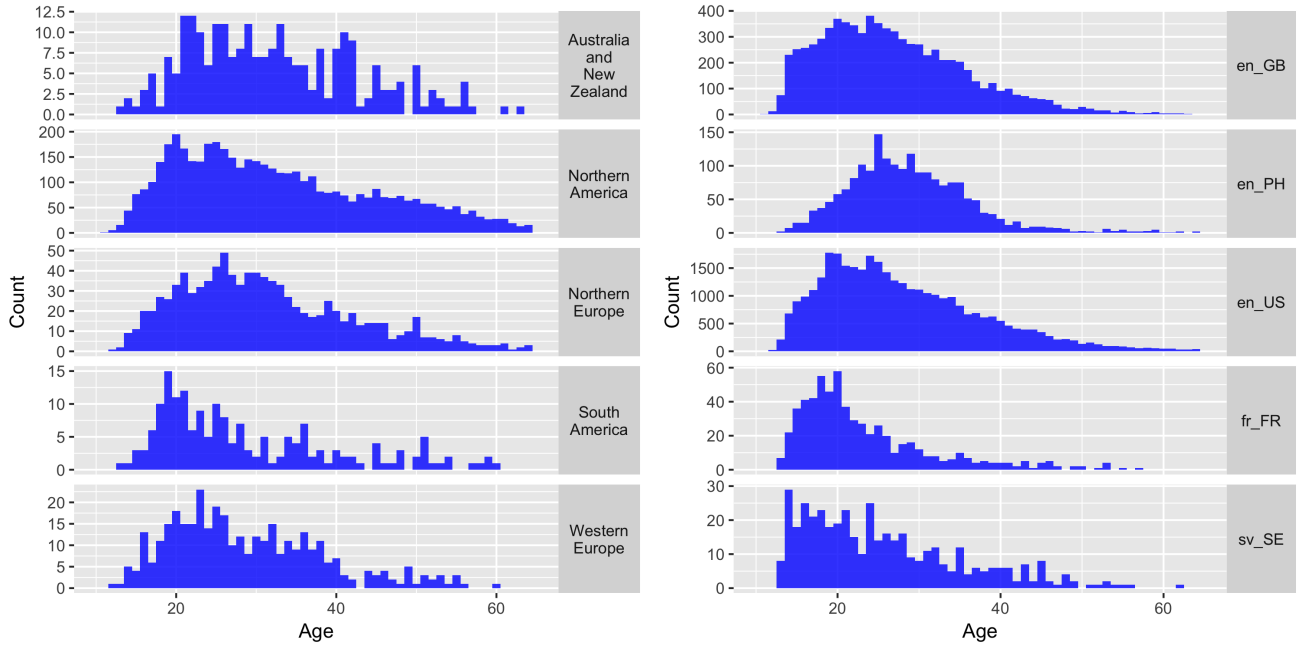

Figure 6: Histogram of age for regions/languages chosen as significantly different distributions by the Wilcoxon test.

8.1.3 Creation Timestamp Distributions by Geographic Attributes

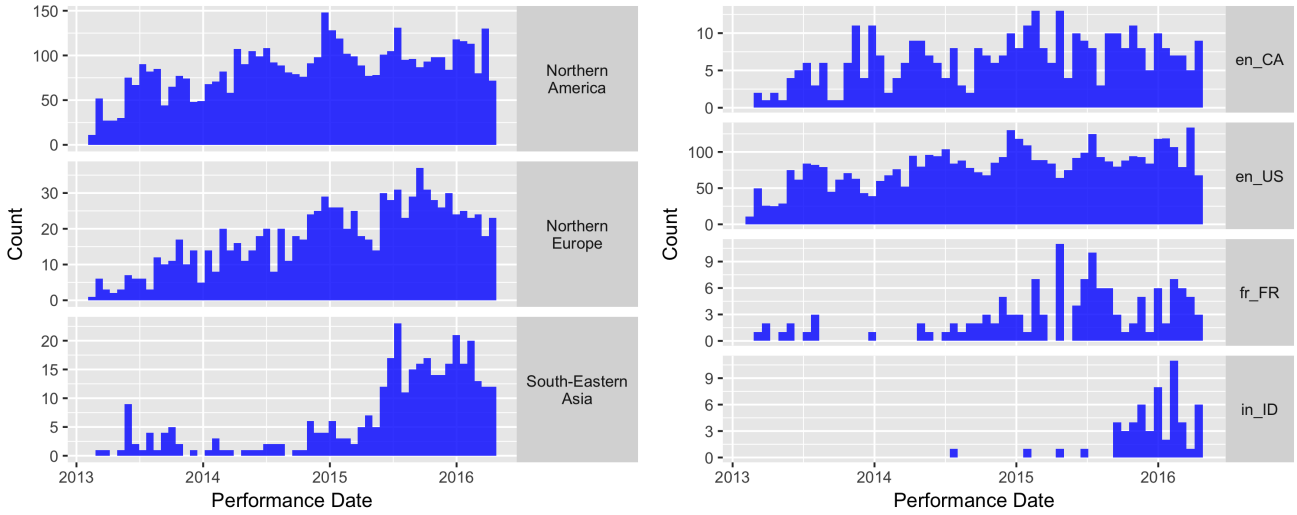

Figure 7: Histogram of creation date for regions/languages chosen as significantly different distributions by the Wilcoxon test.

8.1.4 Creation Timestamp v.s. Age

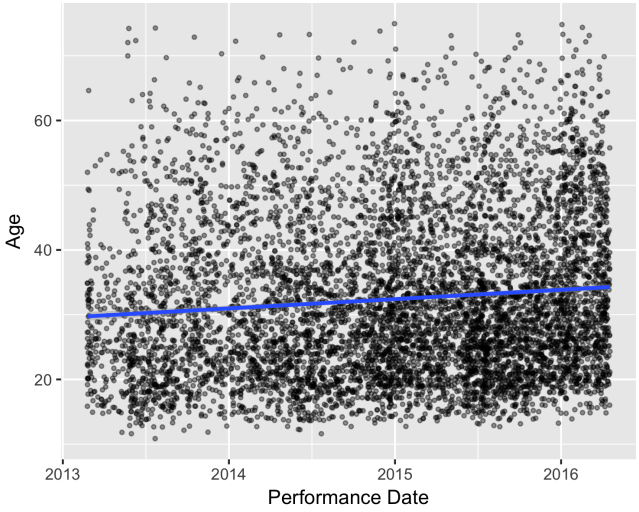

Figure 8: Age graphed against creation date with a best-fit line and 95% confidence band.

8.2 Metadata Figures for Let It Go

8.2.1 Gender Proportions by Geographic Attributes

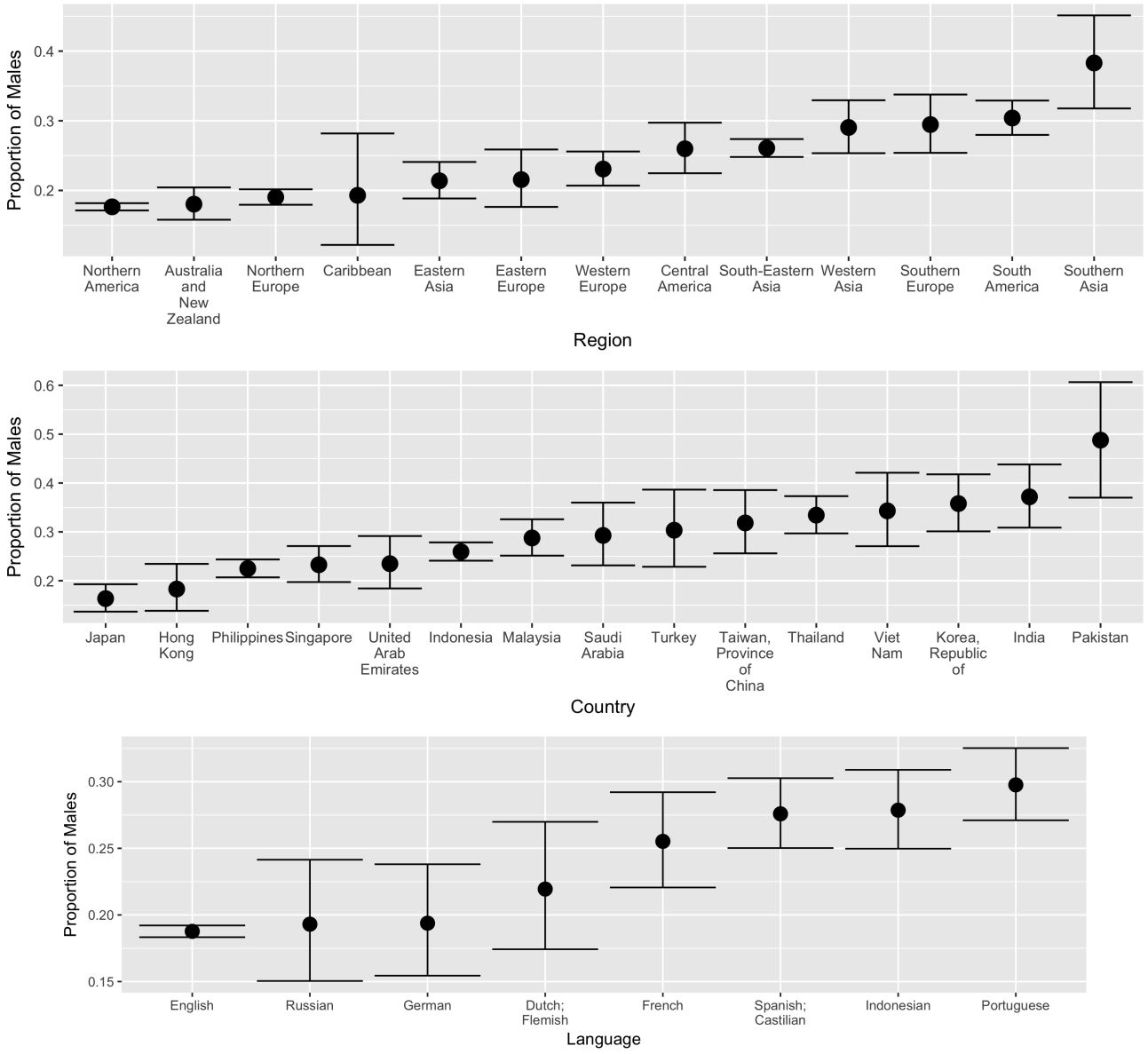

Figure 9: Proportion of singers who are male with 99% confidence bands split by region/country in Asia/language where the number of singers exceeds 70, 150 and 500, respectively.

8.2.2 Age Distributions by Geographic Attributes

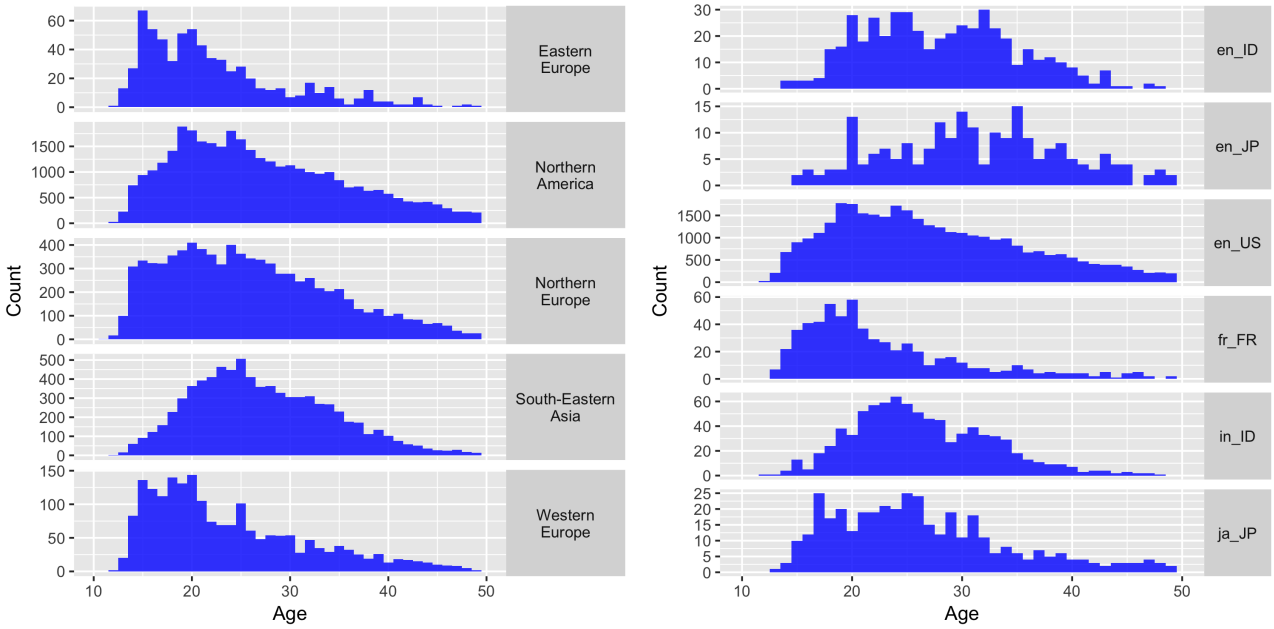

Figure 10: Histogram of age for regions/languages chosen as significantly different distributions by the Wilcoxon test.

8.2.3 Creation Timestamp Distributions by Geographic Attributes

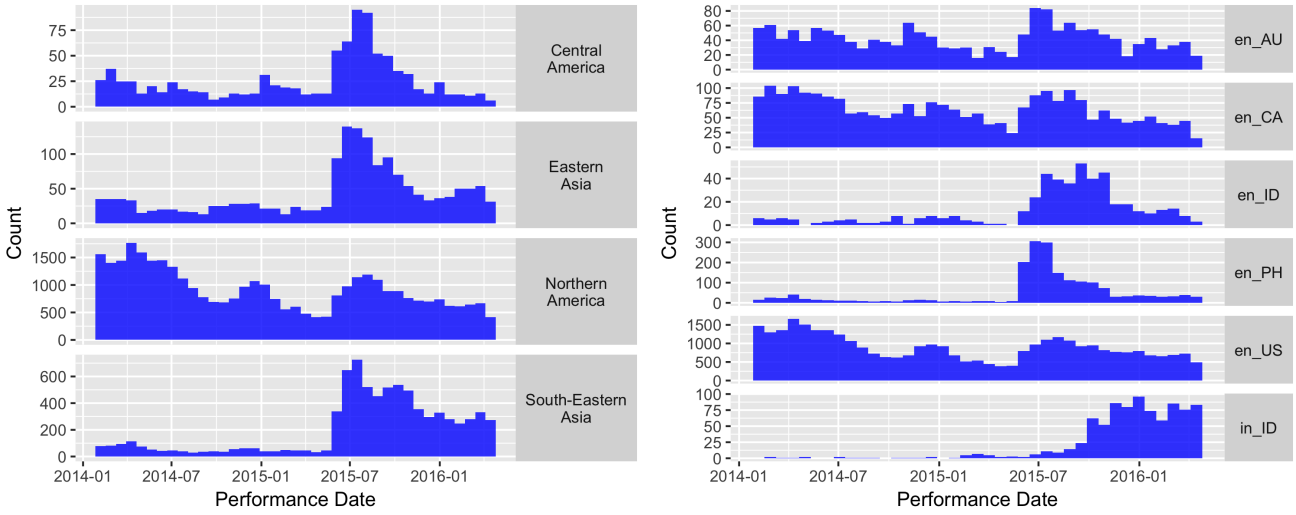

Figure 11: Histogram of creation date with bins of size 3 weeks for regions/languages chosen as significantly different distributions by the Wilcoxon test.

8.2.4 Creation Timestamp v.s. Age

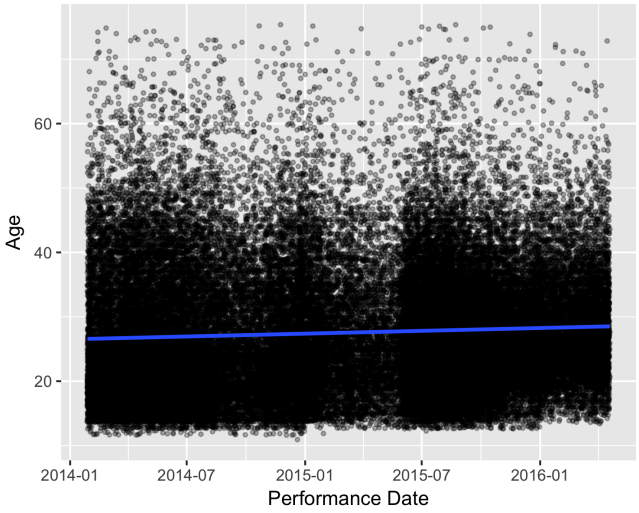

Figure 12: Age graphed against creation date with a best-fit line and 95% confidence band.

8.3 Voice Quality Figures for Amazing Grace

8.3.1 Gender v.s. Voice Quality Metrics

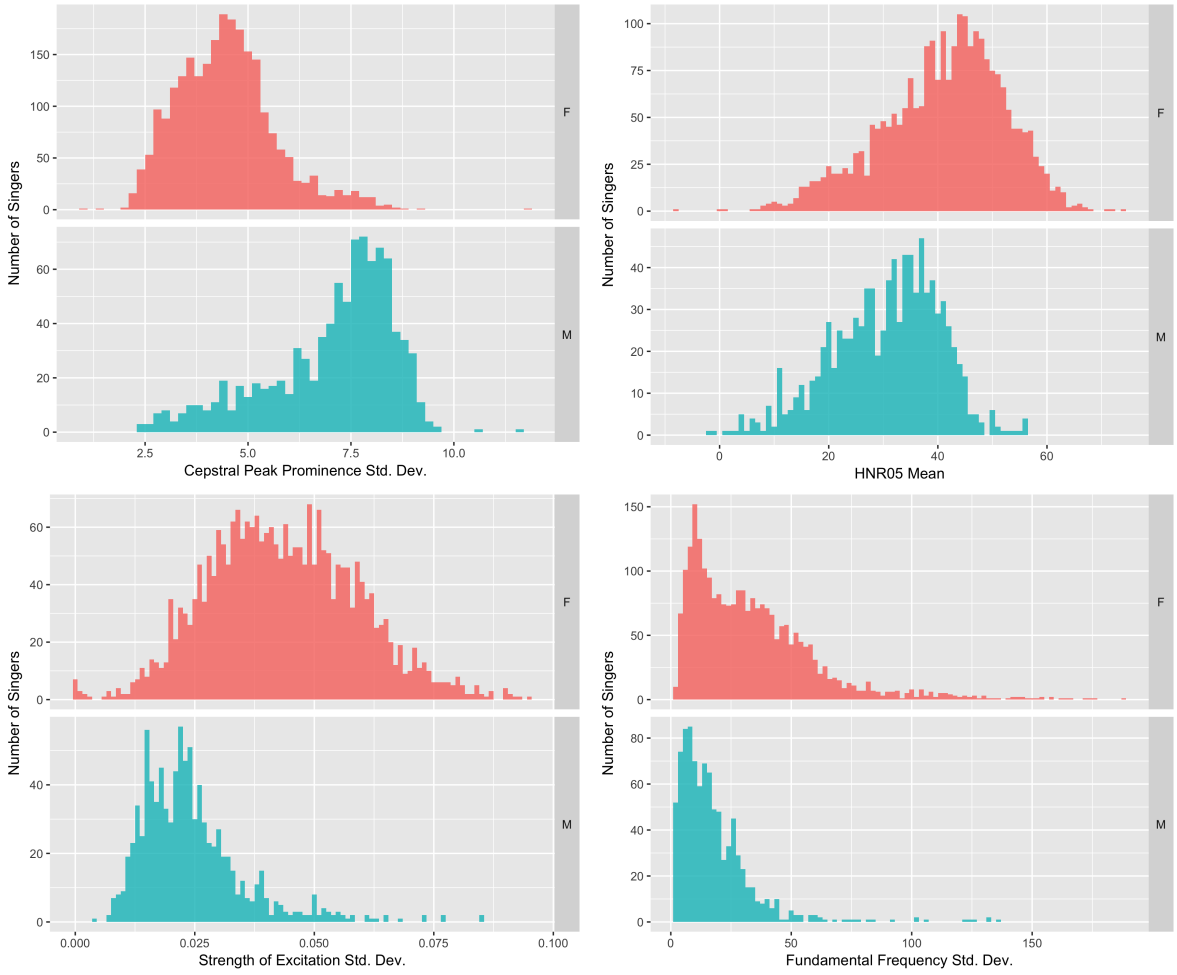

Figure 13: The distribution of some key voice quality metrics in Amazing Grace plotted for each gender.

8.3.2 Age v.s. Voice Quality Metrics

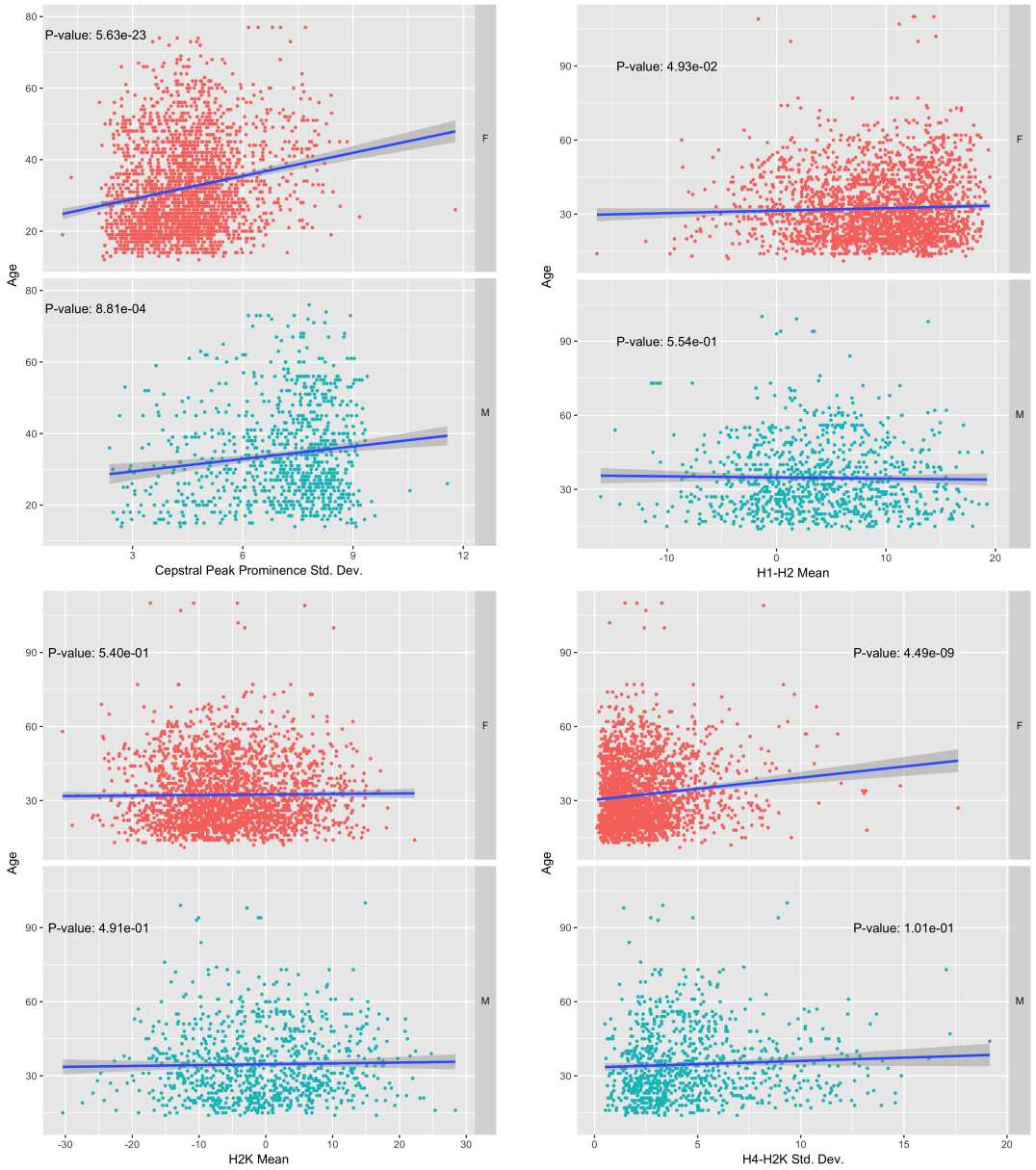

Figure 14: The distribution of some key voice quality metrics in Amazing Grace plotted against age, broken down by gender with a best-fit line and 95% confidence band.

8.4 Voice Quality Figures for Let It Go

8.4.1 Gender v.s. Voice Quality Metrics

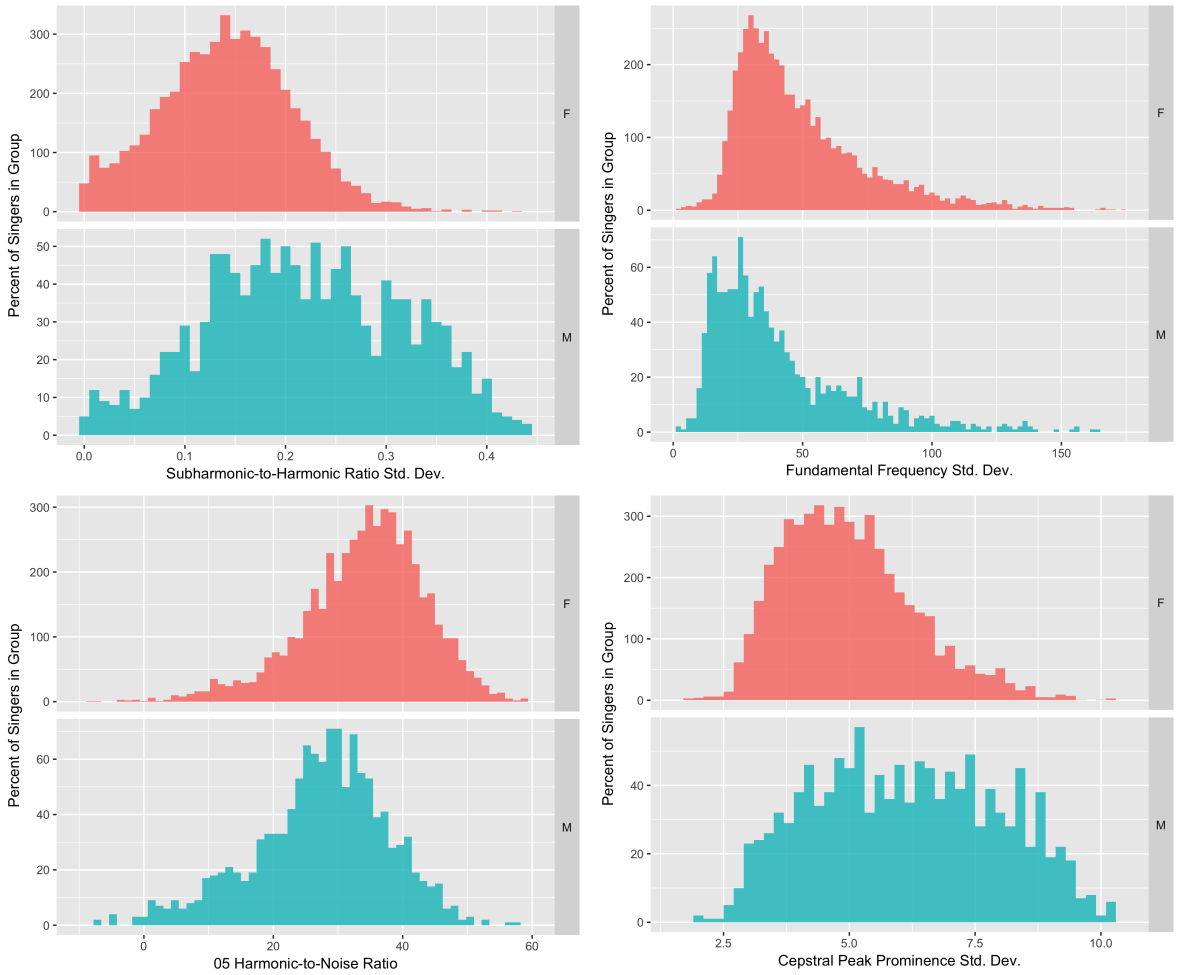

Figure 15: The distribution of some key voice quality metrics in Let It Go plotted for each gender.

8.4.2 Age v.s. Voice Quality Metrics

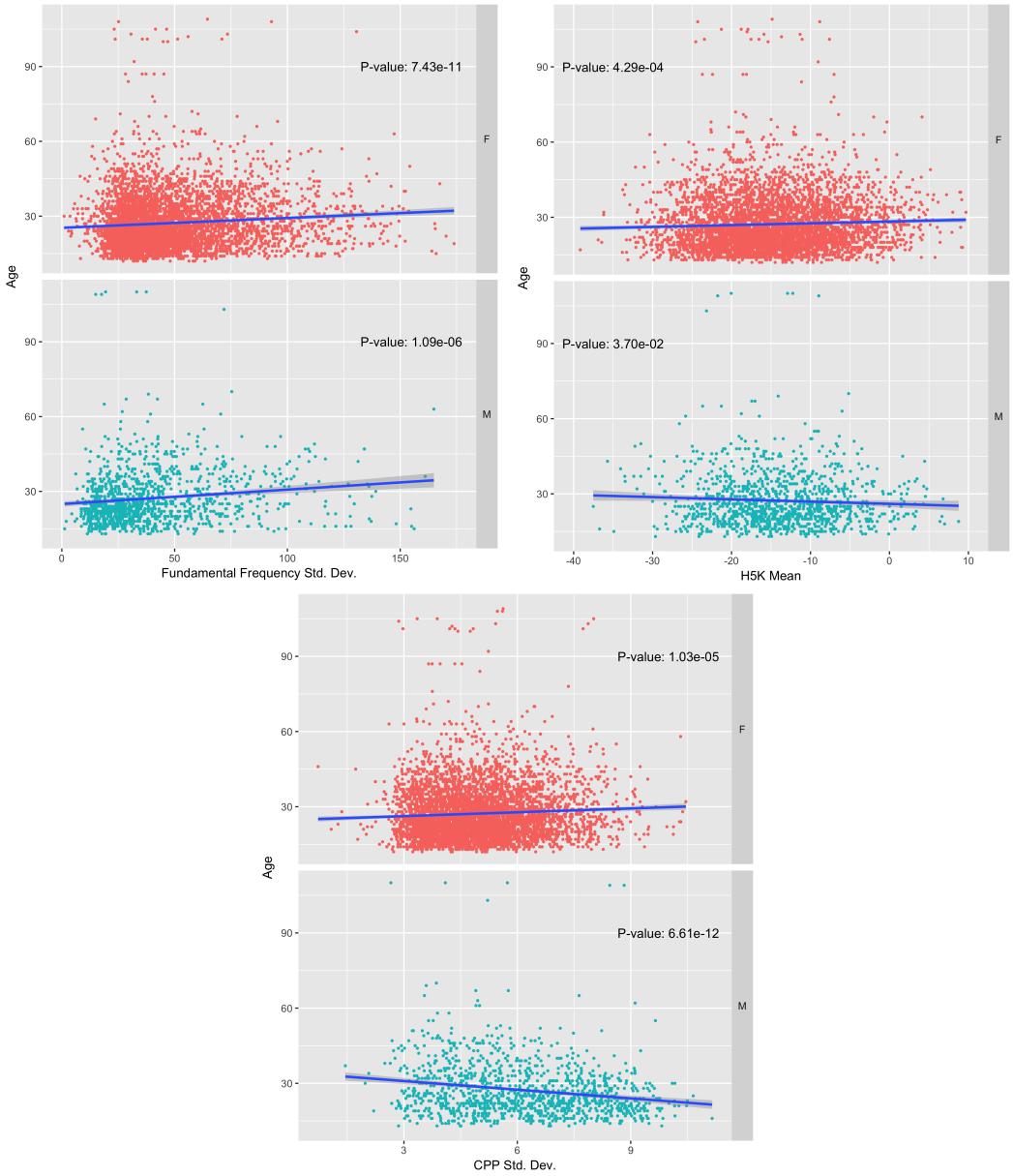

Figure 16: The distribution of some key voice quality metrics in Let It Go plotted against age, broken down by gender with a best-fit line and 95% confidence band.

References

[ABY18] Vincent Arel-Bundock and CJ Yetman. countrycode: Convert country names and country codes. https://cran.r-project.org/web/packages/countrycode/index.html, 2018.

[ASSV17] M. Araya-Salas and G. Smith-Vidaurre. warbler: an r package to streamline analysis of animal acoustic signals, 2017.

[Bec16] Kory Becker. Identifying the gender of a voice using machine learning. http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/, 2016.

[CS18] Perry Cook and Jeffrey C. Smith. Damp: Massive research data sets from mobile music performances, 2018.

[GCA10] Yen-Liang Shue Gang Chen, Xue Feng and Abeer Alwan. On using voice source measures in automatic gender classification of children's speech. *Proc. Interspeech 2010*, pages 673–676, 2010.

[JF18] Rob Tibshirani Noah Simon Balasubramanian Narasimhan Junyang Qian Jerome Friedman, Trevor Hastie. glmnet: Lasso and elastic-net regularized generalized linear models. https://cran.r-project.org/web/packages/glmnet/index.html, 2018.

[Kuh18] Max Kuhn. caret: Classification and regression training. https://cran.r-project.org/web/packages/caret/, 2018.

[LW18] Andy Liaw and Matthew Wiener. randomforest: Breiman and cutler's random forests for classification and regression. https://cran.r-project.org/web/packages/randomForest/index.html, 2018.

[Man17] Usha Manjunatha. Cepstral analysis of voice in young children and adults. *Global Journal of Otolaryngology, Vol 11, Issue 1*, 2017.

[Rid17] Greg Ridgeway. gbm: Generalized boosted regression models. https://cran.r-project.org/web/packages/gbm/index.html, 2017.

[Smi13] Jeffrey C. Smith. Damp: Stanford digital archive of mobile performances. https://ccrma.stanford.edu/damp/, 2013.

[TTR18] B. Atkinson T. Therneau and B. Ripley. rpart: Recursive partitioning and regression trees. https://cran.r-project.org/web/packages/rpart/index.html, 2018.

[YSY11] C. Vicenik. Y.L. Shue, P. Keating and K. Yu. Voicesauce: A program for voice analysis. *Proc. 17th ICPhS Hong Kong*, page 1846–1849, 2011.