Sample of Working with Spectrogram Voice Data

The following is an example of a classification forest model built upon measurements from spectrogram data. The work is done on the dataset voice.csv compiled by Cory Becker and available on Kaggle. This dataset was created using functions from the libraries seewave and tuneR in R's Cran repository to create and analysze spectrograms from a set of voice recordings. Each entry (row) in the data is a voice recording and each variable (column) is a measurement taken off of the spectrogram (excepting the variable "label" which is the biological sex of the entry, reported directly from Becker's data). This dataset was compiled with the express intention of predicting biological sex. My intent, however, is to construct a model that groups entries by the individual who speaks them. It is not wholly unrelated, previous work has shown biological sex to be one of the foremost predicatble features of a voice and integral to voice recognition. I am working on scraping my own data from the VoxForge database which is more suited to my purposes (it contains multiple voice entries from specific individuals). I am using this dataset here with the intention of providing an example of what spectrogram data looks like and how it can be incorporated into a predictive model vis a vis the request in the feedback to my first project proposal.

A spectrogram plots the frequency of a sound wave over time, usually with amplitude represented as a color axis. Here is an example of a spectrogram for one of the audio files I am working with for the full project.

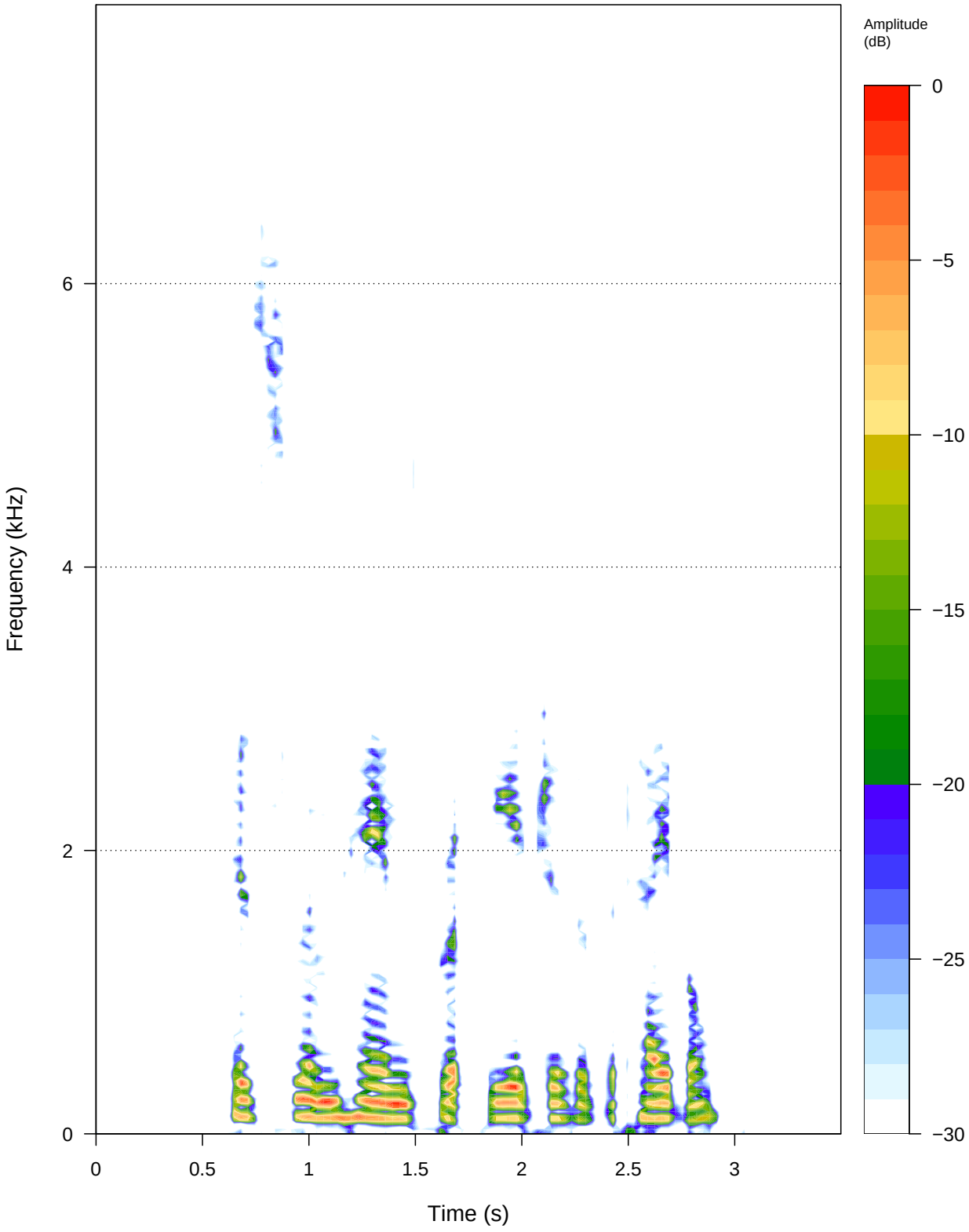

Figure 1: Spectrogram

The majority of variables in this dataset are ordinary summary statistics of the frequency distributions in Becker's spectrograms but there are a few that are more uniquely suited to analyzing voice data. Fundamental frequency, spectral entropy, and spectral flatness in particular tend to be better measures of human vocal characteristics. Other, similarly appropriate variables, that I hope to make use of that are not present here include number of harmonics and amplitudes of the waves.

The variables in the voice dataset are defined as follows:

meanfreq: mean frequency of the wave(in kHz)

sd: standard deviation of frequency of the wave

median: median frequency (in kHz)

Q25: first quantile for frequency (in kHz)

Q75: third quantile for frequency (in kHz)

IQR: interquantile range for frequency (in kHz)

skew: skewness of the spectrogram's frequency, computed as $S = \sum_{i=1}^{N}(freq_i - meanfreq)^3 \times \frac{1}{sd^3}$. S<0 indicates left skew, S>0 indicates right skew, S=0 indicates perfect symmetry

kurt: kurtosis of the spectrogram's frequency, computed according to $K = \sum_{i=1}^{N}(freq_i - meanfreq)^4 \times \frac{1}{sd^4}$, measures the spectrogram relative to the normal curve. K<3 indicates fewer items at center and tails than expected from normal but more in the shoulders, K>3 indicates more items at the center and tails than expected but fewer at the shoulders, and K=3 indicates a perfect normal curve.

sp.ent: spectral entropy. Describes the complexity of a sound wave, ie how much information is being conveyed. A pure synthetic tone has low entropy, a recording from a crowded diner has high entropy. Roughly speaking it indicates how "noise-like" a sound is compared to how "tonelike". Ranges between 0 and 1.

sfm: spectral flatness. White noise produces a spectrogram that looks nearly flat, with only minor rising and falling around its central tone. This measure indicates how steady and near to white noise a spectrogram is. Sfm closer to 1.0 indicates a very monotonic sound and human voices typically score much closer to 0.0.

mode: mode frequency

centroid: frequency centroid. Computed as $C = \sum_{i=1}^{N}(freq_i - meanfreq)^2 \times \frac{1}{sd^2}$

peakf: peak frequency (frequency with highest energy)

meanfun: average of fundamental frequency measured across acoustic signal. Fundamental frequency is the lowest frequency produced by oscillation of the object. In terms of hearing, it is the lowest pitch or tone that you hear with harmonics rising above it where the the frequency of the sound waves exactly double that frequency. The mean fundamental frequency thus gives an approximation of a person's most comfortable "natural" pitch, though it can be forced higher or lower by modulation of the voice.

minfun: minimum fundamental frequency measured across acoustic signal

maxfun: maximum fundamental frequency measured across acoustic signal

meandom: average of dominant frequency measured across acoustic signal. Dominant frequencies are local maximums of the frequencies, the apexes in the wave.

mindom: minimum of dominant frequency measured across acoustic signal

maxdom: maximum of dominant frequency measured across acoustic signal

dfrange: range of dominant frequency measured across acoustic signal

modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range.

label: male or female

Sample Regression Tree

Split the data into training and testing sets.

```
voice = read.csv("/home/class19/gyoung19/StatComps/voice.csv")
```

```
set.seed(36) #for reproducability
train <- voice %>%
  sample_frac(0.80) %>%
  mutate(label = as.factor(label))

test <- voice %>%
  setdiff(train) %>%
  mutate(label = as.factor(label))

x_train <- model.matrix(label ~ ., train)[, -1]
x_test <- model.matrix(label ~ ., test)[, -1]

y_train <- train %>%
  dplyr::select(label)

y_test <- test %>%
  dplyr::select(label)
```

Create random forest on training set, using default params of 500 trees considering 4 variables at each step with no maximum tree length.

```
set.seed(2250) #reproducability
rf <- randomForest(label ~ . , data = train, importance = TRUE) #produce random forest model
rf
```

```
##
## Call:
##  randomForest(formula = label ~ ., data = train, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 2.17%
## Confusion matrix:
##        female male class.error
## female   1247   24  0.01888277
## male       31 1232  0.02454473
```

```
estimate1 <- predict(rf, newdata = test) #assess accuracy of model on "new" data
misclass1 <- ifelse(estimate1 != test$label,1,0) #create vector indicating where misclassifications occ
mean(misclass1) #get percentage of misclassifications on test set
```

```
## [1] 0.009463722
```

Our model seems to be able to differentiate between the biological sex of the speakers with a high degree of accuracy, misclassifying only 1.26% of the time. We would also like to check which of these many variables were most informative to our model.

```
importance(rf) #list variables and importance
```

```
##             female      male MeanDecreaseAccuracy MeanDecreaseGini
## meanfreq 10.223607 11.481962             14.99099        26.770622
```

```
## sd       16.851184 15.945382          21.81419        97.635865
## median    9.696393 14.214570          16.44871        18.682786
## Q25      18.925172 23.620311          28.38364       159.398463
## Q75      10.617579 12.139675          15.66817        13.699193
## IQR      20.394572 35.714711          38.58051       244.137512
## skew     10.372807  8.376263          12.99566        13.602288
## kurt     10.526120  8.171386          12.32354        10.229845
## sp.ent   14.865760  9.826768          17.64559        57.016189
## sfm      16.982912 12.861330          20.18108        41.084571
## mode     10.639303 11.233951          13.37967        19.905143
## centroid  9.520364 12.591789          16.01013        22.352289
## meanfun  42.814560 65.873813          71.07732       473.254813
## minfun   11.732828 12.073388          16.17091        12.338545
## maxfun    9.386261  7.286574          11.16917         6.341982
## meandom  12.123758  8.913113          13.98629         9.560521
## mindom    5.866225 10.972492          11.43918         8.878321
## maxdom   13.384097  9.478714          15.51142        12.034714
## dfrange  15.439257  9.604017          17.23617        11.119503
## modindx  14.207147  7.524171          14.92663         8.412018
```

It seems that the mean fundamental frequency leads by quite a large margin and probably merits further consideration in my continuing work. This is not at all unexpected, fundamental frequency has been useful in differentiating biological sex of a speaker before. Reports on the actual numbers vary, but generally state that the average range of male fundamental frequency is somewhere between 85 to 180 Hz and the average female range is between 165 and 255 Hz.