Applied Data Science — Speech Recognition

Portland Data Science Group – June 23-30, 2019

Data Summary: Over the course of four weekends, the group worked in small groups to build algorithms that identify which words are spoken in short segments of audio. A training set of 678 short audio samples was compiled. Each sample represents a spoken number, “zero” through “twelve”, for a total of thirteen possible classifications of each audio sample. Each audio sample potentially contained background noise or other audio artifacts.

I. Remove artifacts

  A. Use GAM to smooth absolute amplitude, producing peaks

  B. Get center and width of highest peak

II. Fit (weighted) Fast-Fourier-Transform (FFT)

  A. Use peak center to translate waveform so that time zero corresponds to peak

  B. Use center and width to obtain case weights (based on Gaussian distance from peak)

  C. Use width to determine length parameter of FFT

  D. Transform centered data to absolute amplitude (instead of signed amplitude)

  E. Get Fourier coefficients (15 harmonics seems sufficient for this problem)

III. Build Naive Bayes Classifier

  A. Get multivariate normal stats for each (true) number spoken

  B. Form the prediction of any set of FFT coefficients as the posterior probability of class assignment, with the stats in III.A forming the prior distribution. (Note: each number is considered equally probable in the prior distribution).

Error rate on training data (from 10-fold cross-validation): 9.5%, assigning each sample the label corresponding to the maximum posterior probability. **Note that the entropy of the posterior distribution also predicts whether the sample will be correctly classified (see Figure 2 below)**.

Figure 1: The clustering heatmap below illustrates the (cross-validated) posterior probabilities. Rows are individual samples, with the left annotation strip showing (by different color) the true number being spoken. The columns represent the possible classifications assigned by the classifier (numbers zero through twelve). The cells of the heatmap show the posterior probability of being classified as the label indicated at the bottom of the columns (intense blue = 100%, white = 0%).

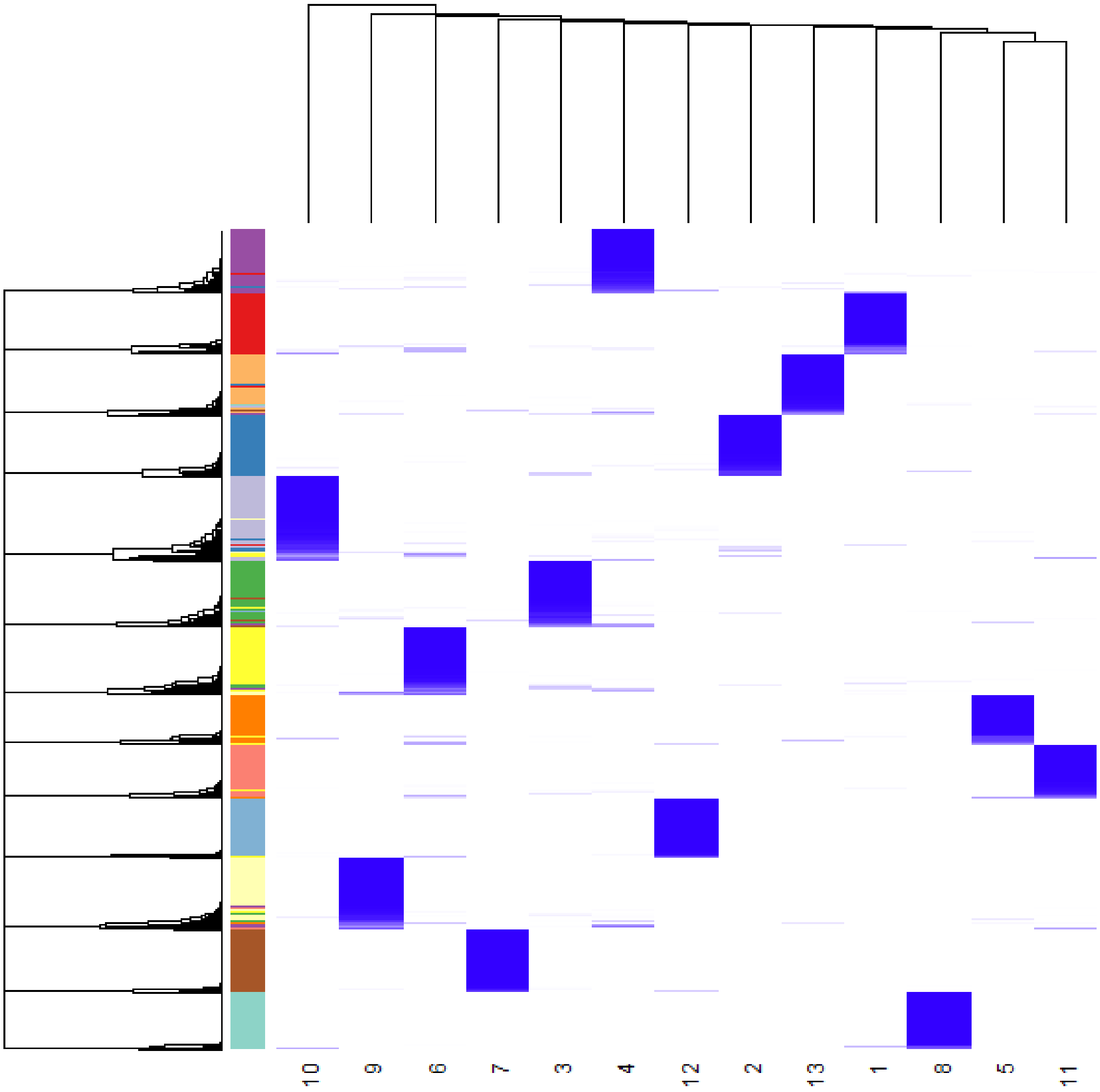


Figure 2: The box-and-whisker diagram below illustrates how posterior distribution entropy predicts correct classification (Wilcoxon $p < 10^{-15}$)

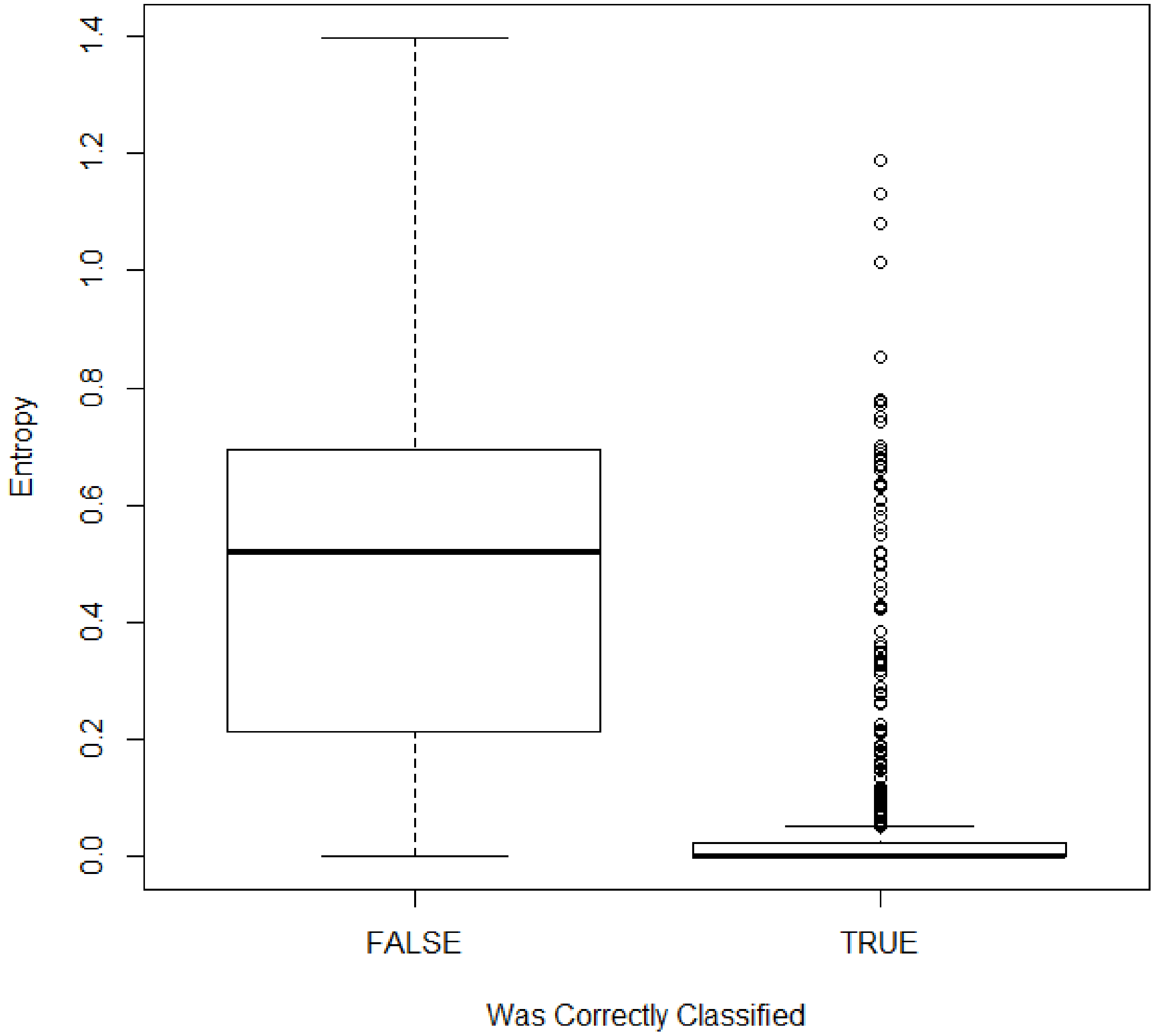


Figure 3: Examples of I - II: raw data (absolute value), smoothed envelopes, and Gaussian weight function for four select samples. Black lines show the absolute value of the raw wave form (0.0625 seconds clipped from front and back). The red solid lines show the envelope calculated using a generalized additive model (mgcv:::gam in R). The blue dotted lines illustrate the shape of the weight function used in calculating the Fourier coefficients. The horizontal green line shows the "width" calculated using the peak height and the curvature of the GAM envelope at the peak. Note that each waveform is centered at its maximum peak, so that the origin time represents the maximum amplitude.

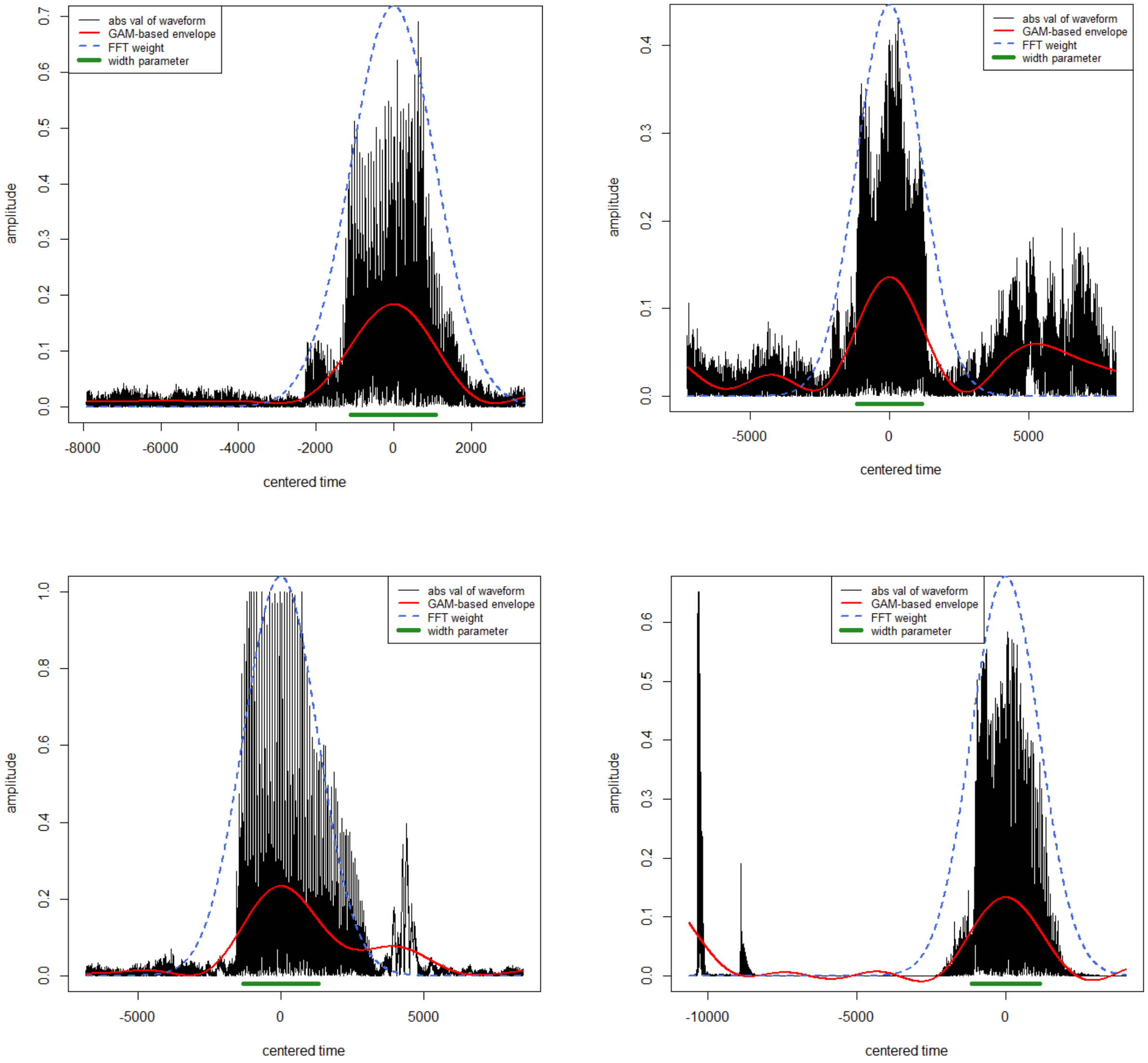


Figure 4: Examples of II - III: scatter plots of select pairs of Fourier coefficients. Color indicates correct classification (which number was spoken).

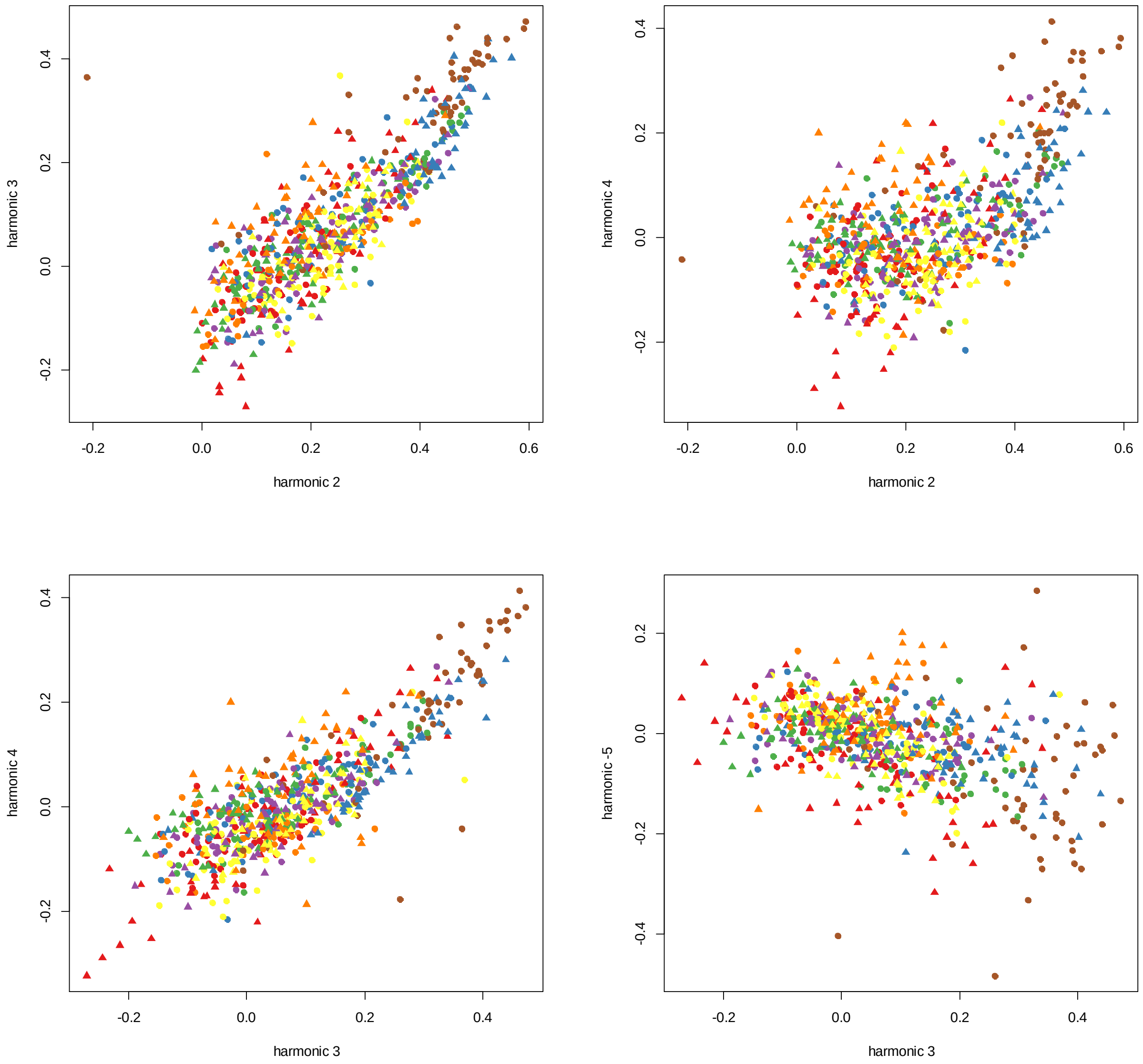


Figure 5: Examples of II - III: Multivariate normal distribution estimated from each individual class, select pairs of Fourier coefficients. These correspond to the scatter plots in Figure 4. Color indicates correct classification (which number was spoken). Each ellipse represents a 95% probability bound for the multivariate distribution implied by the data.

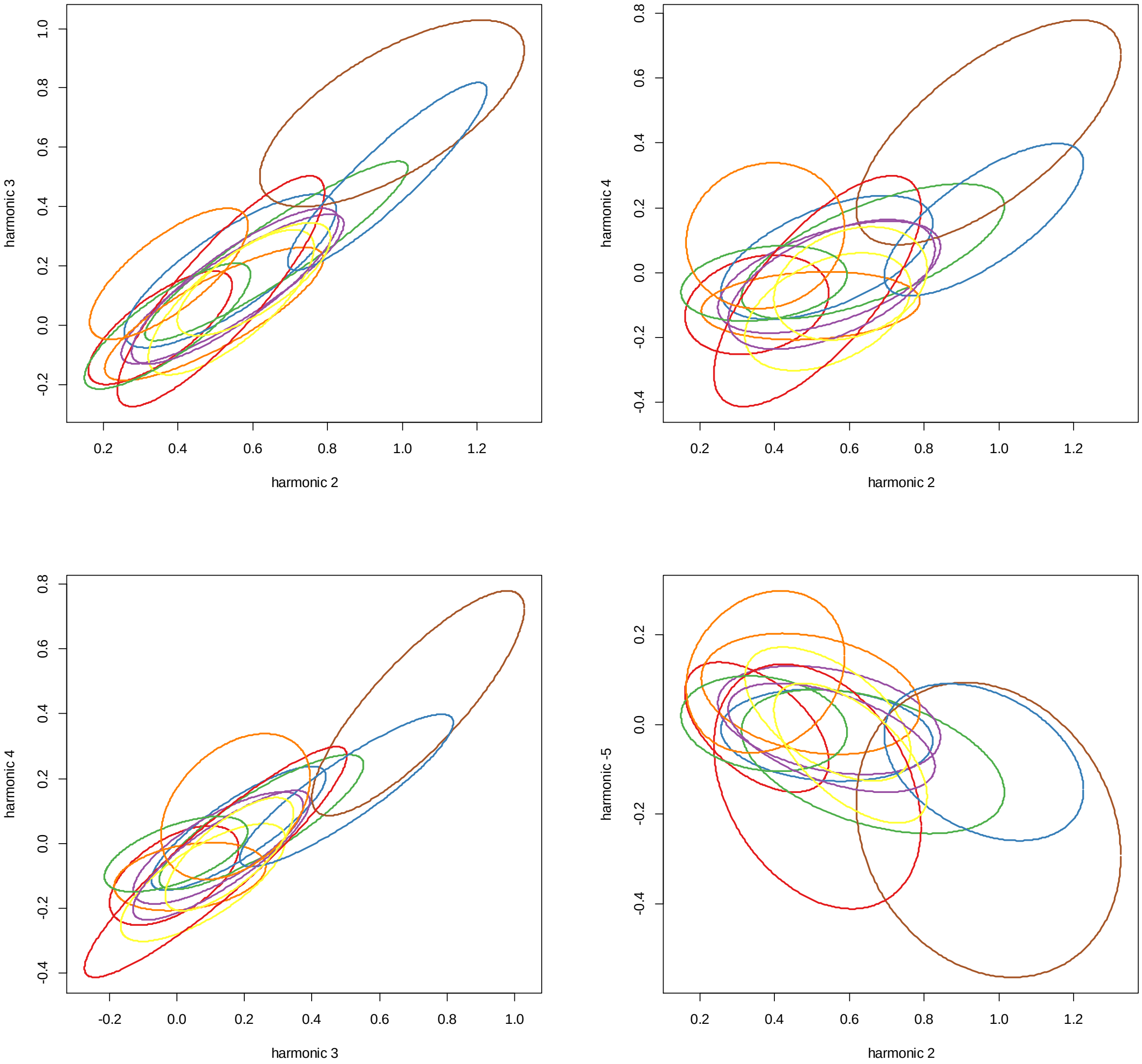
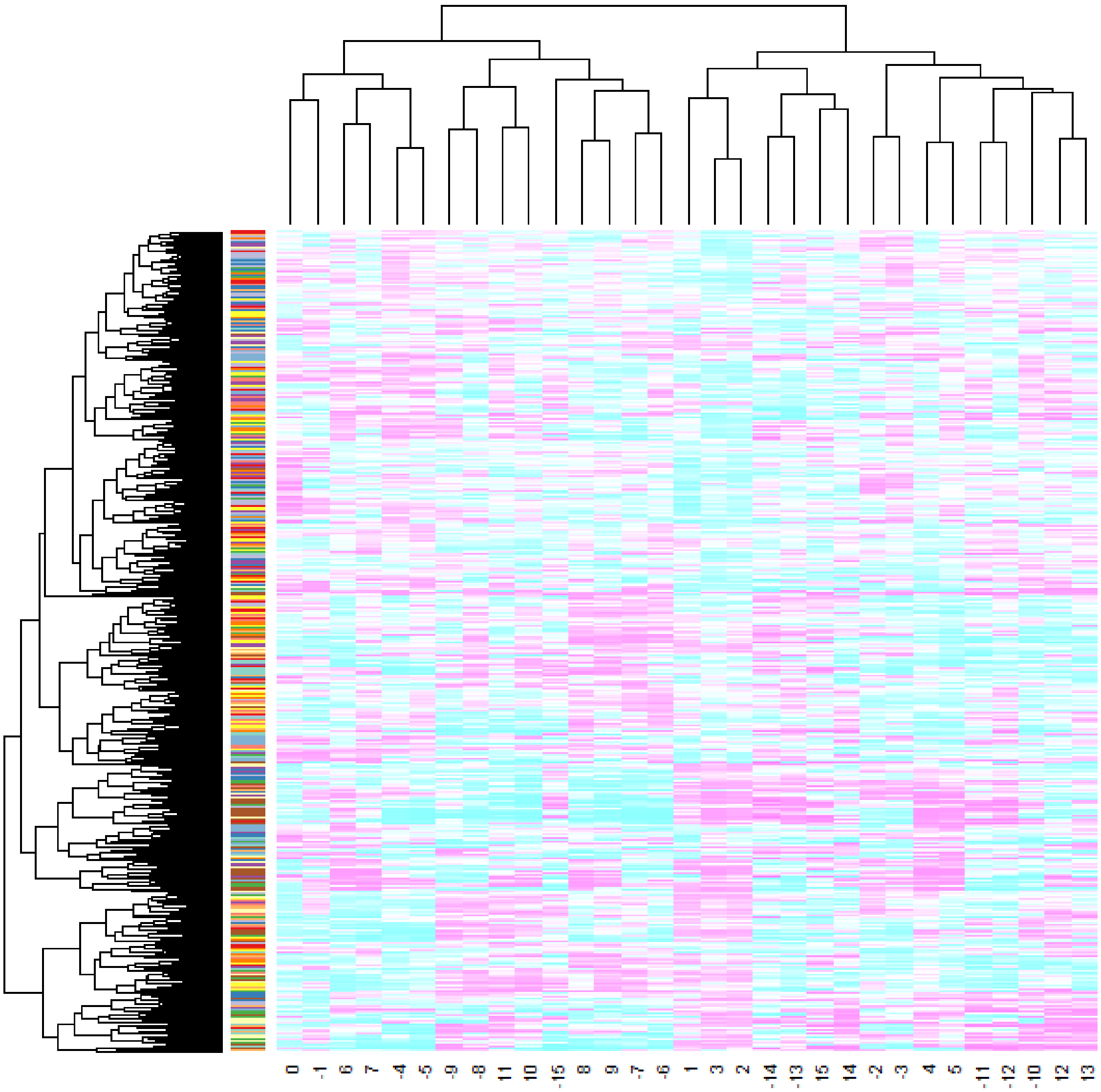


Figure 6: Examples of II - III: Clustering heatmap of all Fourier coefficients. Rows are individual samples, with the left annotation strip showing (by different color) the true number being spoken. Each column represents a different harmonic coefficient. The cell color indicates magnitude within column (pink = high, turquoise=low).


0
-1
6
7
-4
-5
-9
-8
11
10
-15
8
9
-7
-6
1
3
2
-14
-13
15
14
-2
-3
4
5
-11
-12
-10
12
13