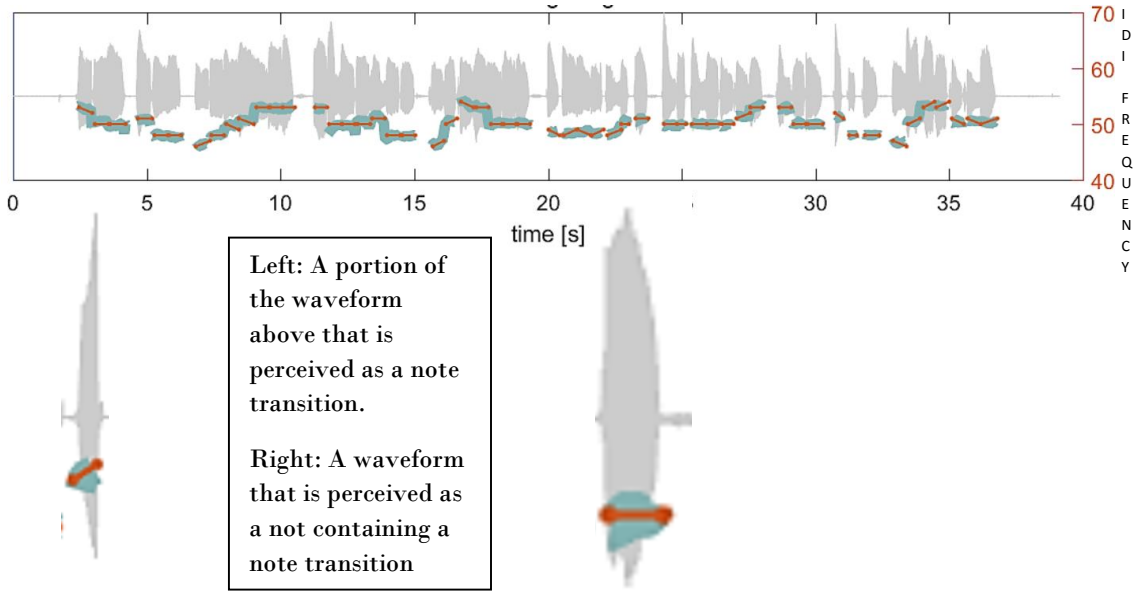
Automatic Rhythm Notation from Single Voice Audio Sources

Jack O'Reilly, Shashwat Udit

The Data & The Challenge:

- The singe human voice is subject to many variations.
- No absolute physical or mathematical definition for note onset with a human voice.
- The standard we are trying to match is what a skilled human listener would transcribe if they were listening to a performance.
- The data input for are algorithm is the waveform itself.
- Several databases of monophonic melodies.
- Two databases which had human annotations to train dataset
- One dataset which was synthesized from the actual notation to provide highly accurate training data.



Left: A portion of the waveform above that is perceived as a note transition.

Right: A waveform that is perceived as a not containing a note transition

Results & Future Directions:

- The second and third portions work reasonably well conditional
  - Accurate Onset and Offset Times
  - Similar styles as the training set
- Possible commercial program may need machine learning algorithm to match style of music
- Primary objective is remaining time is to improve detection of note onset and offset times.
- GMM approach
  - Train two GMMs -- one for audio snippets with onsets and one without.
  - Use Expectation Maximization Algorithms to train
  - Develop detection function that evaluates probabilities from each of the GMMs and compares samples

Methods:

- The problem we're attacking has three parts:
  - detecting note onsets and offset times through SVMs,
  - taking the raw differences and converting them into intended note lengths using k-means clustering
  - Use of multinomial Gaussian combined with tempo to turn note length into note type.
- Classification approach: classify section of audio signal as 'containing' or 'not containing' an onset
- Failed SVM kernels included:
  - Linear
  - Polynomial
  - Radial basis
- Result for all three: overly 'pessimistic' models that always report the naively more likely option -- no onset
- Feature vector of Mel-Frequency Cepstral Coefficients
- Conventional speech / voice processing technique, often used for pitch estimation and other applications
- Attempts to represent features of voice signals that are intuitively significant from a psychoacoustic standpoint
- Maps energy to log frequency scale
- Feature vector was 69 coefficients: 23 cepstral coefficients, and the instantaneous first and second derivatives of each said coefficient.
- (also attempted with 13 coefficients for a feature length of 39)