The Beatbox system is comprised of three main components:

- The DSP unit accepts voice input, then cleans and analyzes the incoming signal.
- The Pattern Recognition subsystem uses frequency characteristics to probabilistically determine the most likely match for the input data.
- The Demonstration system is a GUI that controls the audio and visual feedback given to the user.

**Digital Signal Processing Unit**

When a user’s voice input triggers the underlying engine, it is converted into a digital signal and passed onto the DSP unit. The signal is divided into tiny windows (~25 ms long) and multiplied by the Hanning window, before its FFT is taken. Our system is modeled as a filter bank like the human ear, allowing for compression of information contained in the signals. Further redundancies are eliminated through Cepstral analysis before handing over the processed signal to the Pattern Recognition subsystem.

**Pattern Recognition Unit**

**Training**: The recognition/classification system is based on the theory of Hidden Markov Models (HMMs). Given the observation sequence, we infer the most likely underlying “hidden state” sequence using the Viterbi algorithm. We then iteratively estimate the parameters of the HMM till convergence is achieved.

**Testing**: Each dictionary sound has a pre-trained HMM corresponding to it. The signal is passed through each HMM; the input is classified in favor of the HMM with the highest estimated likelihood.

**Motivation**

Speech recognition is a key tool in the design of the next-generation user-friendly computer application. A major obstacle remaining in the way of this goal is the detection of stop consonants, sounds created by stopping the flow of air in the mouth and letting it go into a burst (ex. b, d, g, k, p & t). Telling stop consonants apart is a difficult problem due to their similarity. Building a system to distinguish stop consonants may help bring speech recognition one step closer to reality.

**Authors**: Priyadarshini Routh, Aryeh Levine, Raphael Levy

**Advisor**: Prof. Lawrence Saul

**Demonstration**

90% accuracy on dictionary sounds if trained on the same user.

80% accuracy if pre-existing training set is used.

**How the System Hears You!**

Waveforms and Spectrograms of /k/, /pf/, /t/