Vowel Recognition using Formant Analysis

Mauro Zabala, Sujay Tadwalkar, and Steven Tsai

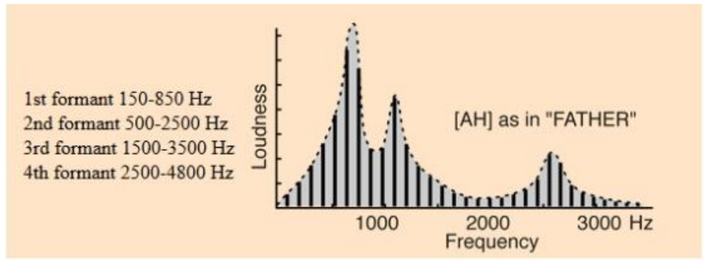
Rice University, Electrical and Computer Engineering, Houston, TX

Introduction

Identifying vowels:

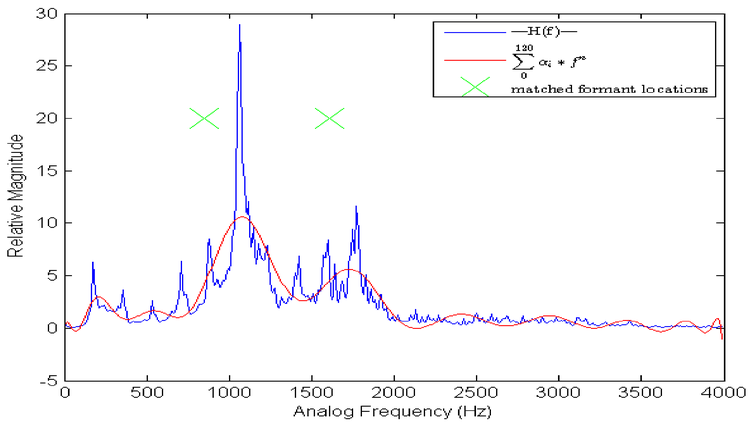
1. Observe the frequency response of the word.
2. Model the frequency response of the system as having all poles so the model has little dependence on the original input signal.

$$\underbrace{x[n]}_{\text{white noise}} \longrightarrow \boxed{H} \longrightarrow \underbrace{y[n]}_{\text{vocal tract}} \qquad \underbrace{\qquad}_{\text{observed speech}}$$

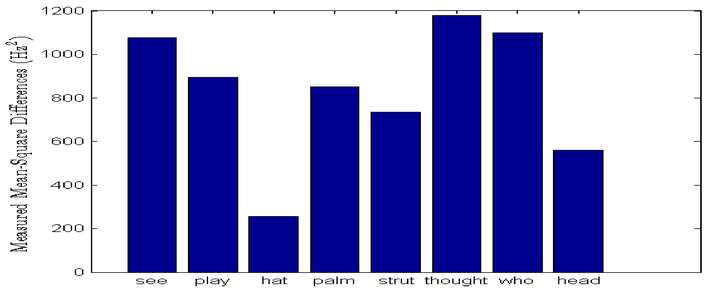


Formants of an 'Ah' sound

One Vowel Recognition

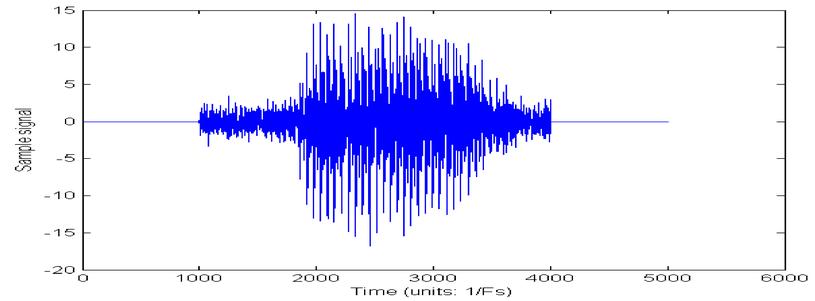


Formant determination of the signal "hat"



Measured mean-squared differences of the signal "hat"

Multiple Vowel Recognition

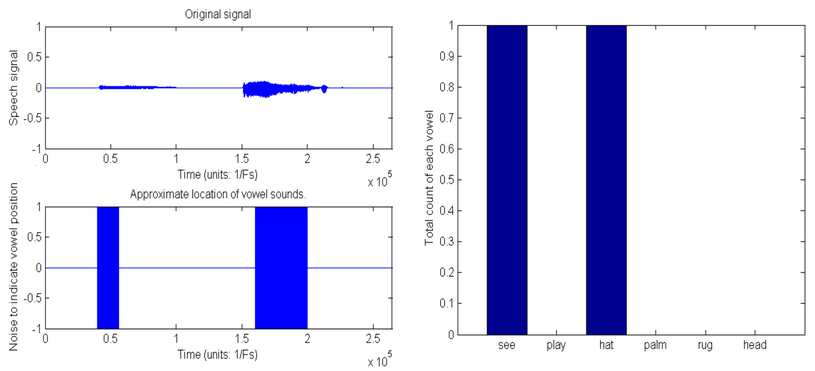


Sample signal of the word "hat" with problematic noise

How to determine the vowels:

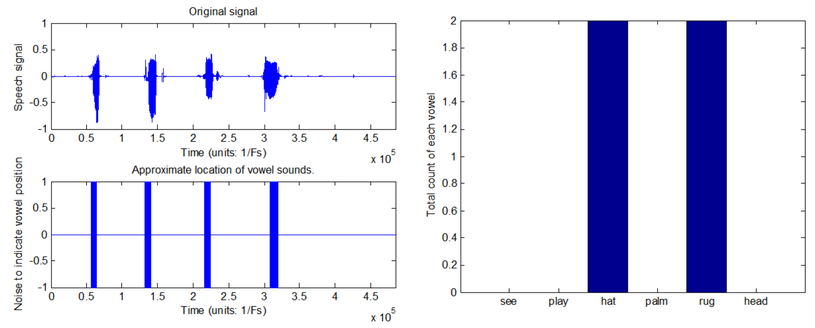
1. Split data into non-overlapping chunks.
2. Preprocess data by detrending and passing through a Butterworth LPF.
3. Identify and compare formant pairs for each chunk.
4. Process formants to identify vowel duration and position.

Dylan Jones says "Bee Baa(pronounced like bat)"



The vowel you said from:
.9075 to 1.2699 seconds is the central vowel of "see"
3.6281 to 4.5352 seconds is the central vowel of "hat"

Nathan Bucki says "Hat Cat Rug Thug"



The vowel you said from:
1.2699 to 1.4513 seconds is the central vowel of "hat"
2.9932 to 3.1746 seconds is the central vowel of "hat"
4.8980 to 5.0794 seconds is the central vowel of "rug"
6.9841 to 7.2563 seconds is the central vowel of "rug"

Auto-Regressive Model

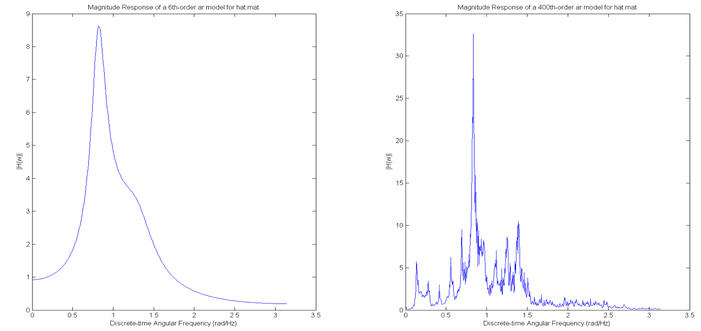
Speech recognition utilizes probability theory to effectively understand the vocal tract system. For iid, w.s.s white noise passed through an LSI system:

$$y[n] = \sum_{k=-\infty}^{\infty} x[n-k]h[k] \quad \Rightarrow \quad S_y[\omega] = \sigma^2 H[\omega]H[\omega]^*$$

We solve for the transfer function with the known system below modeling the output as a recursive average. The recursive average system can be referred to as the autoregressive model.

$$y[n] = x[n] + \sum_i \alpha_i\, y[n-i] \quad \Rightarrow \quad Y_t = \epsilon_t + \sum_i \alpha_i Y_{t-i}$$

To optimize the model of the transfer function, we compared different orders of the AR model in terms of accuracy and precision.



Low vs High Order of the Autoregressive Model

A combination of a high order AR model and polynomial regression most accurately approximates the transfer function.

Limitations

1. Requires very clear enunciation. This is difficult to establish because humans naturally change their pitch at the start and end of words.
2. User must say each vowel for a moderate duration of time. Wavering one's voice would affect the success of the program.
3. The program is calibrated to average formant values across all ages and genders. Accents and gender (low vs. high pitch) may affect accuracy of the results.

Acknowledgements

We would like to thank Jeff Lievense, Richard Baraniuk, and Vivek Boominathan for providing guidance and instruction on the project.