**Part 1: Feature Extraction**

1. **What audio features did you choose to extract and why?**

* Mel Cepstral Coefficients(MFCC’s) and LPC coefficients are selected as features from audio speech of child and parent, because sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out.
* The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. LPC help to add pitch information to the features.

1. **Are there any features you intentionally didn't use?**

* I ignored frequencies 0 to 600Hz during MFCC’s feature extraction, since addition of background noise causes phase shifts across (70 – 1000Hz).
* I ignored silence data in a signal to avoid misinterpretation of signal energy in a frame while extracting features(MFCC’s).
* Length of a signal, since it is recommended to train and test audio signals with constant length since energy of signal in each frequency bin if signal length vary.

1. **What steps did you take to transform the original .wav files to prepare for feature extraction?**

Steps:

* Separated child and parent speech in a wav files offline using Audacity tool, to label features/ visualize the difference between features.
* Silence removal in audio file using Voice Activity Detection(VAD) algorithm. (Note: Currently done this offline while parent and child speech separated from audio file)
* Child and parent audio files are created separately for input to feature extraction.
* Convert raw PCM data to float audio range {-1 to 1} samples.
* Background noise suppression (Note: currently ignored background noise effecting frequencies from signal while feature extraction, In future it need adaptive noise cancellation or beam forming techniques as per the hardware design to be robust for real world environment)
* Divide complete signal in to 30ms frames (for stationary behavior), to maintain stationary nature of speech signal.

1. **How well do you think this audio reflects real-world conditions? What types of sounds might be harder to handle, and what edge cases would you prioritize detecting or filtering first?**

* Real world conditions have other types of noises like (TV, Car, cafeteria, PUB, etc.,) noise suppression algorithms need to be tuned and tested under these environments.

* Multiple speakers talking at a time will be a challenging task. I prioritize background noise and double talk condition handling solutions by using beam-forming and adaptive noise suppression techniques.
* Parent speaking to child (assume Microphone is with child) at distance >3 meters is difficult to extract features. need calibration for such cases were desired signals with low energy.

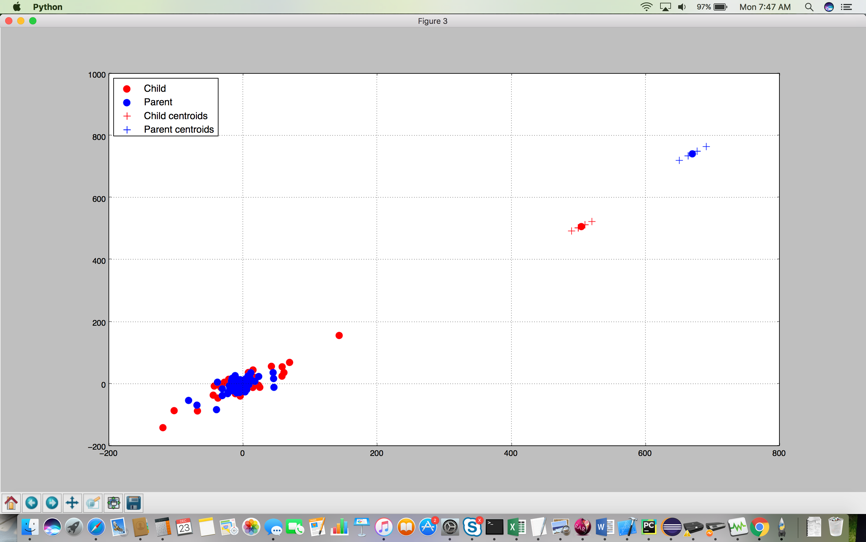
1. **Is there anything you would want to add, optimize, or improve if you had more time?**

* As a child has short vocal cords, they produce short air waves and consequently a high pitched voice. As a child grows, the vocal cords become longer. This variation need to be considered and personalize our device to child age growing from 0 to 3years. ML models need to be trained with data with respect to age (ex: 3monts, 6months, 1year, 2year baby).
* Background noise cancellation and ML Algorithms need to be fine tuned for other real world noises{cafeteria, pub, park, car, etc.,) simulation.
* Child voice have lot of other sounds utterance than words this need to be considered before feature extraction.
* Need a preprocessing unit to separate child noises like (Uuuuu.., Aaaaa, etc.,) to avoid miss interpretation of features/frequencies present in a signal.

* **I** would have added preprocessing units for background noise suppression and VAD algorithm to ignore silence in speech signal to make extracted features more informative.
* More data need to be collected for generalizing child features and preprocess for training ML model to make system robust.
* Fine tune feature extraction algorithm parameters to meet our child voice feature extraction
* Optimize the code to reduce MIPS and help running it on low power device.

1. **If you were to annotate these audio files to train a supervised learning model, what labels would you include? Assume each row of the dataset represents a 10-second audio snippet, and provide an example column header.**

* MFCC coefficients at filter banks {1,…n}, LPC coefficients, signal length will be the main features to be considered as label.



* As shown in figure from MFCC coefficients Codebooks are generated using LBG algorithms for visibility about feature distribution. Centroids of Child and parent are generated from codebook.
* From above observation in figure there is a distinct in vocal tract from mother to child, this is evaluated by offline processing input data like silence removal, ignoring Background noise effecting frequencies.

# Part 2: Modeling

## Task

Your challenge is to build a model to detect the gender of a voice. Always predicting "male" will achieve 50% accuracy. Our naive model based strictly on average dominant frequency achieves accuracy of 61% on training data and 59% on test data. We expect you can do significantly better.

1. **Determine which properties are statistically significant for determining gender. Report your results.**

* All the given properties in input data are statistically significant, as it helps to estimate person vocal information.
* Extracting MFCC and LPC coefficients from speech sample and applying statistics to convert it as a features help to improve system performance to be robust for real world application.

1. **Build a full logistic regression model and evaluate its performance. Report your results, including accuracy on the training and test set.**

Full logistic regression model implemented and Evaluated in python, please refer “Logistic\_regression” project in repo. Observations are as given below

Train Accuracy: 78%

Test Accuracy: 77.6%

1. **Train any model of your choice to achieve an accuracy above 80%.**

Using Tenforflow I could able to achieve below performance, please refer to “Gender\_prediction” project in repo. Observations are as given below.

Train Accuracy: 94%

Test Accuracy: 89%

**Questions**

1. **What changes did you make in Section 3 to increase the accuracy of your ML model? Include a description of features you selected or created, parameters you tuned, and a discussion of tradeoffs.**

* To improve accuracy of prediction Tensor flow framework with Keras layers API configured as per the application feasibility, In our application for gender identification with given features of 21 for 3168 male/female voices.
* Hidden are layers sequenced as dense layers with activation functions as shown below.
  + tf.keras.layers.Dense(20, activation=**'relu'**),  
    tf.keras.layers.Dense(10, activation=**'relu'**),  
    tf.keras.layers.Dense(1, activation=**'sigmoid'**)
* Tried other activation functions like “tanh” for hidden layers, as there is no significant improvement in accuracy switched to ‘Relu’ for taking advantage of less computations.
* Number of neurons fine tuned for achieving optimal performance in minimal neurons to reduce computations. Activation function ‘relu’ will help on fast convergence.
* In Dense layer each neuron is connected with every other neuron on neighbor layer help to improve accuracy for our application as statistical features about the voice are related to each other.

1. **Is there anything you would want to add, optimize, or improve if you had more time?**

* MFCC and LPC coefficients as a feature give more accurate information about human vocal tract.
* Fine tune ML algorithm to achieve optimum performance with less computational complexity and higher convergence speed.
* Collect and train more data with different real time conditions and ambient noise environments, collect data of male and female from different regions to make system robust to all region people.
* Background noises like car, cafeteria, PUB, etc., simulation and suppression for being robust to real world data.
* Silence removal algorithm to make extracted features more information about person voice.

1. **How might the model you built perform in the real world? What technical or ethical considerations would you weigh when deciding whether to deploy such a system?**

* Model currently built is a base model which require parameter tuning like epoch, layers and preprocessing to be robust to real world data.
* Accuracy may come down with above model when it uses in real world noises.
* Privacy of data collected for processing application should be a priority before deployment.
* Choose right MCU, DSP to meet algorithm requirements like MIPS, RAM, ROM, to avoid algorithm delays and underruns during processing of audio signal.
* Need to enable multi microphone techniques to run adaptive noise cancellation using adaptive algorithms (NLMS/RLS/APA, etc.,) for improving performance of system.
* Before deployment of system it need to be tested in a anechoic chamber with noise simulation setup, to confirm system is robust for real world noises.

1. W**hat would you do differently if you were trying to detect adult vs. child speech instead of male vs. female voices?**

* As a child has short vocal cords, they produce short air waves and consequently a high pitched voice. As a child grows, the vocal cords become longer. This variation need to be considered and personalize our device to child age growing from 0 to 3years. ML models need to be trained with data with respect to age (ex: 3monts, 6months, 1year, 2year baby).
* Child voice have lot of other sounds utterance than words this need to be considered before feature extraction.
* Need a preprocessing unit to separate child noises like (Uuuuu.., Aaaaa, etc.,) to avoid miss interpretation of features/frequencies present in a signal.
* Since in current solution it considered statistics of all frequencies in signal, there will be miss interpretation when system is in real world environment with low SNR, In this case system will interpret noise frequencies also as a features of a speaker voice.
* More frequency and pitch features required to be collected to estimate their vocal tract, MFCC and LPC coefficients are considered as features. This help to analyze more frequency ranges than in given solution.
* While tuning the system in anechoic chamber before deployment, addition to the real world noises (TV, Car, cafeteria, PUB, etc.,) for device like mobile phones, smart speaker etc., have to add other background noises specific to child, like playing with music toys, playing with objects, kitchen noise, etc.,