Pronunciation Assistance Based on Automatic Speech and Facial Recognition

Dr. Maria Pantoja
Computer Engineering
Santa Clara University

Marie Bertola
Modern Language
Santa Clara University

Pronunciation Improvement

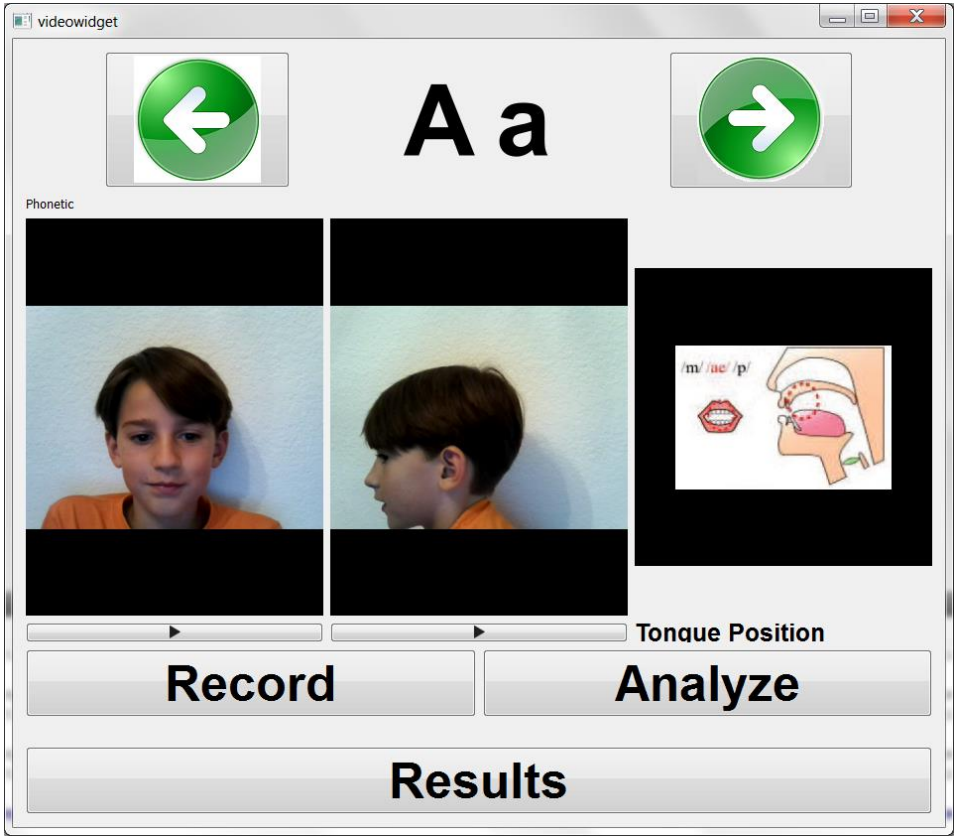
L2 learning assistance instructional tool.

Assessing student's pronunciation

Providing accurate corrective feed-back.

The model presented integrates speech and image recognition technology.

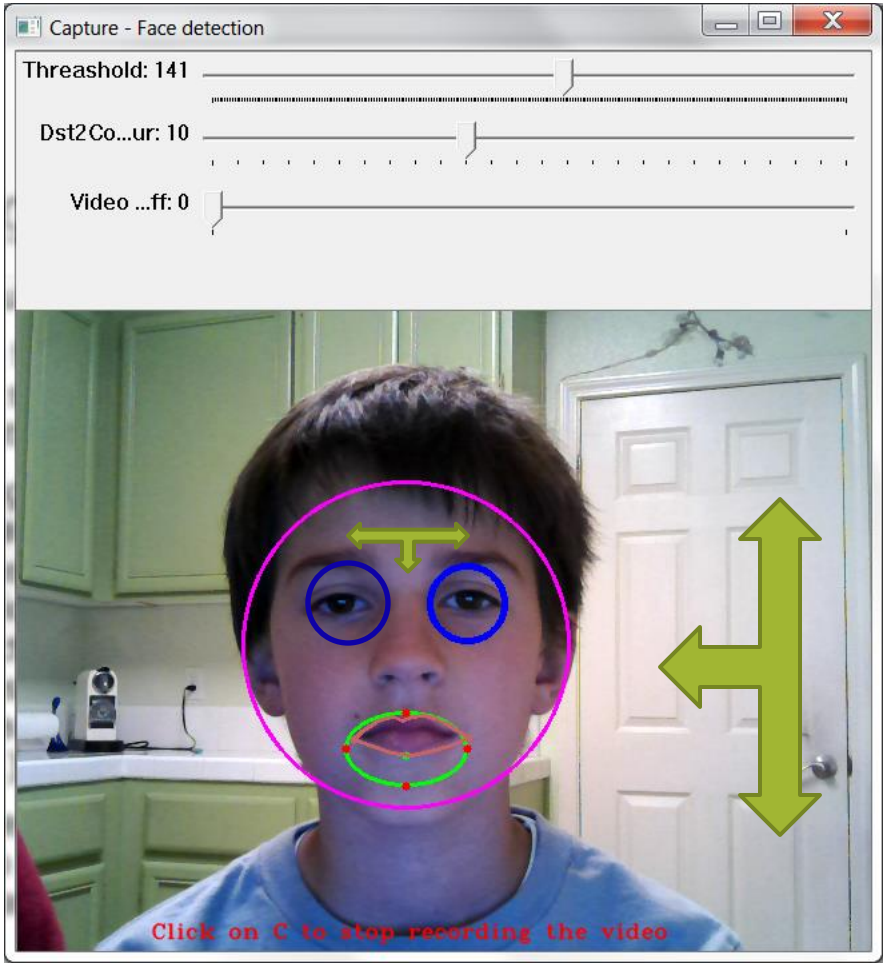
Providing feed-back and data to evaluate the model's performance.



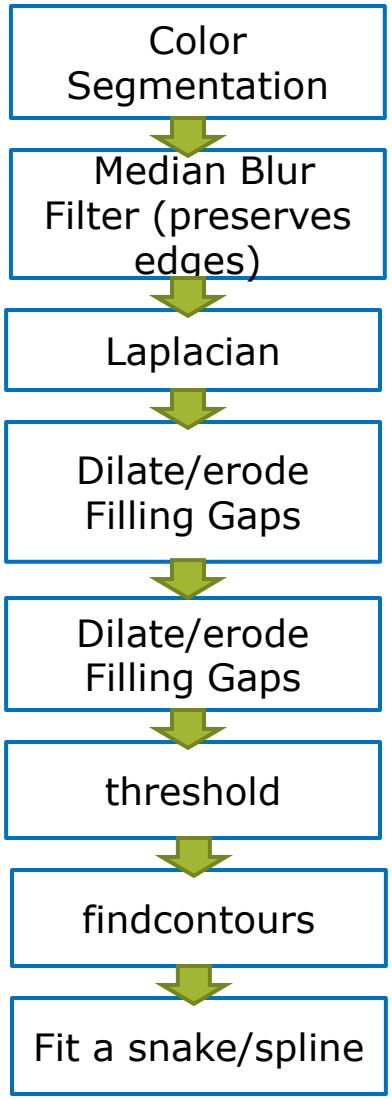
Lip Segmentation

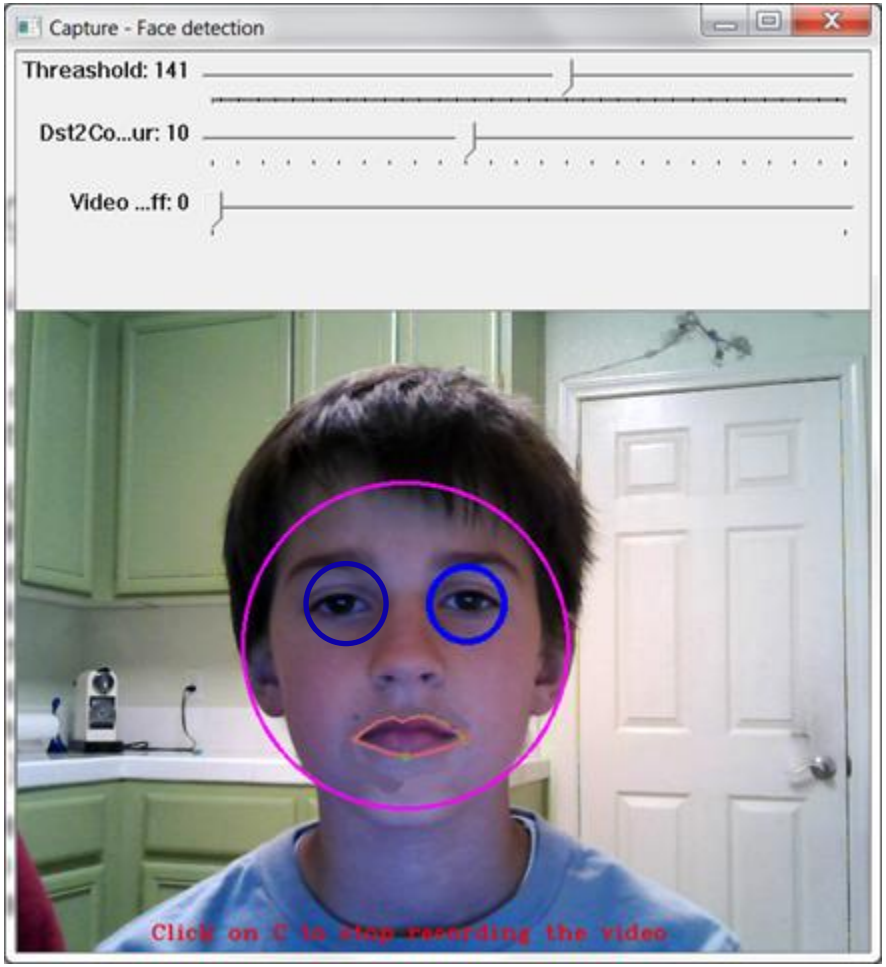
Location of Mouth:

- User will be front facing the camera so it relatively easy to locate the face.
- We are using OpenCV face detector Feature-based Cascade Classifier (Paul Viola and Michael J. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. IEEE CVPR, 2001)
- Using simple human face geometry to restrict the possible location of the mouth
  - 1/3 lower face
  - Between the eyes



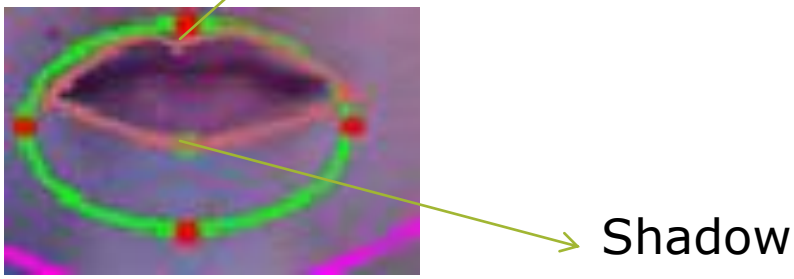
Lip Segmentation





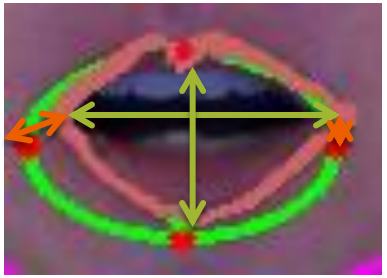
Pseudo Code for Lip Segmentation

- Lip segmentation using color only creates a very noisy image
  - Red is prevalent in both skin and lips
  - The difference between Red and Green values is higher on lips
- Luminance changes. Light sources usually comes from above the speaker
  - The top lip contour is illuminated, the bottom contour is in shadow
- Mouth geometry



Lip Movement tracking

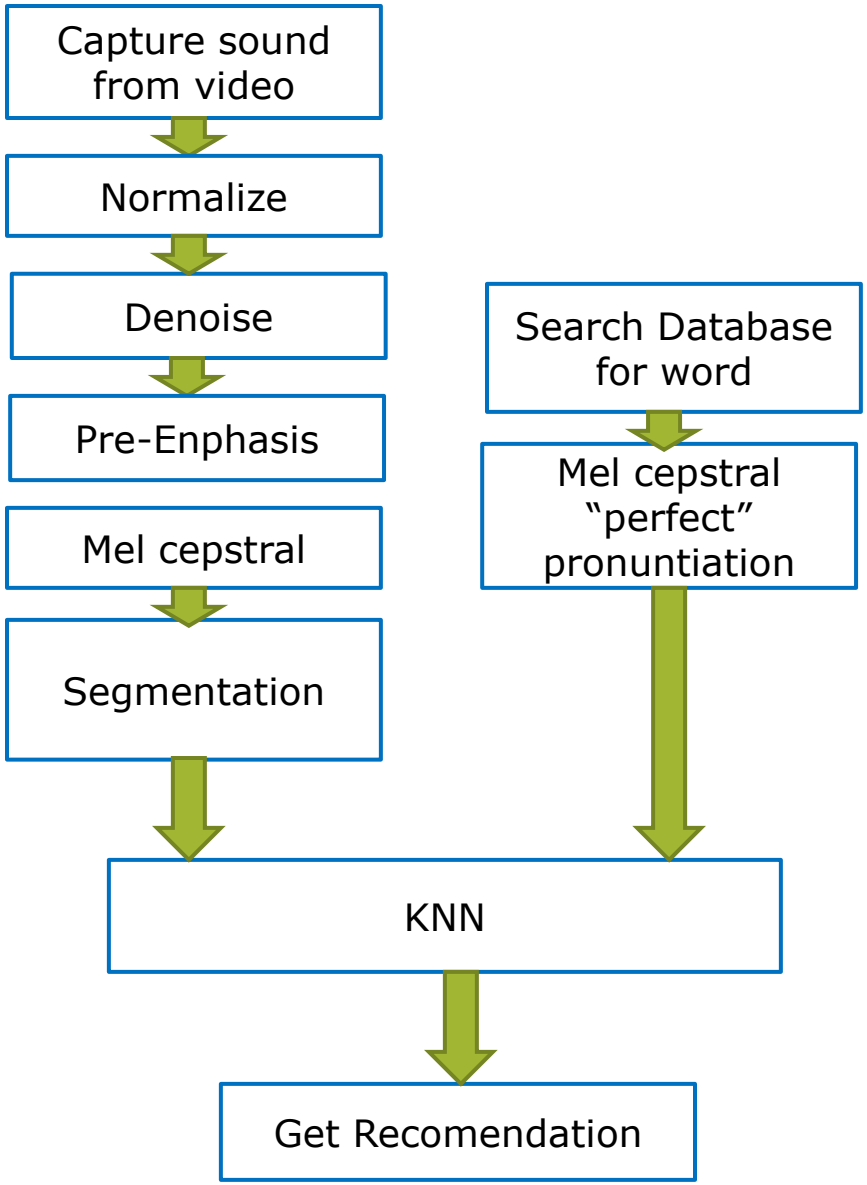
- Keypoints tracking
- Frame by frame we recalculate the 4 pints.
- Problems on the corners when mouth is wide open
  - Restrict search based on the mouth contour
  - Get minimal intensity for the contour



lip segmentation CPU vs GPU

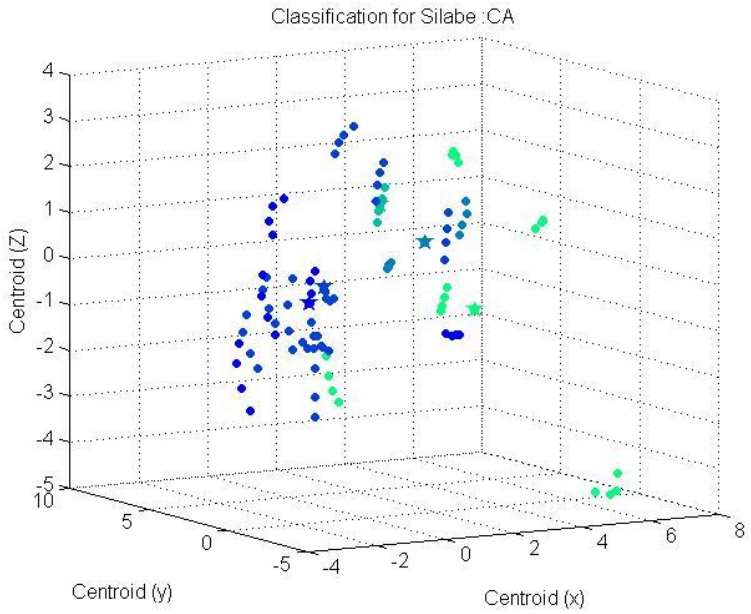
	CPU	GPU
Face Detection		
Mouth Detection		
Erode/Delate		
Blur Filter		
Spline		

Audio analysis



K-NN

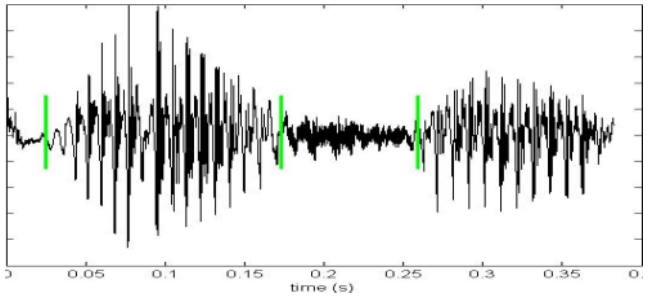
- Supervised classification algorithm
- Training Phase: Storing the feature vectors (MFCC coefficients) and class labels of the training samples
- The centroids for the different pronunciation rules are calculated (represented as starts)
- Classification Step: the user audio is “classified” (assign a recommendation for correcting pronunciation) by assigning which is most frequent among the k training samples

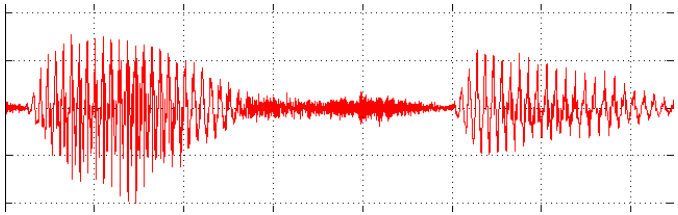


Total Samples: 70
Pro. Rules:
short a
long a
american a
wrong vowel
wrong consonant
totally wrong
Number of samples identified correctly: 62

Audio Rate

- Compare rate for each of the letters in the word
- Recommended things likes:
  - Make the a longer/Shorter
  - Etc

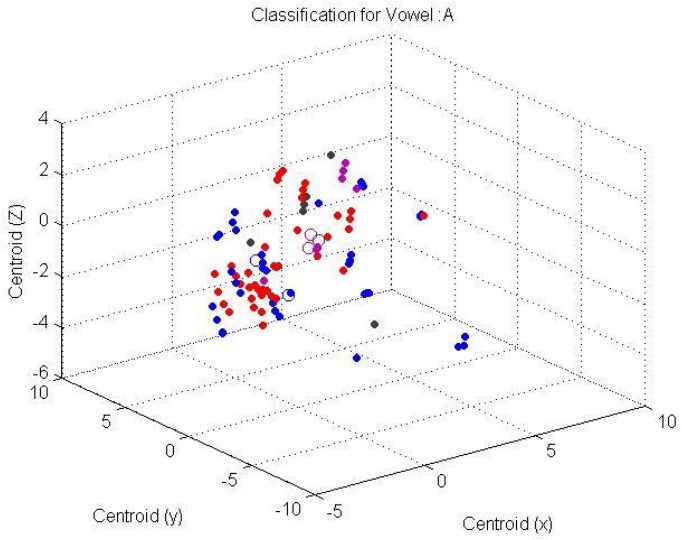


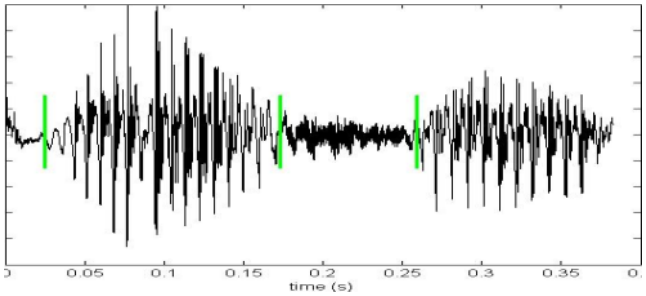


Results Audio Vowels :A

- Train User first with just vowels .
- Vowels have a stronger signal and
- Also only 5 vowels in Italian
- Classification is easy :
  - Wrong vowel
  - Length too short/long
- High Success rate
  - 2 errors in 70 samples

    (due to background noise)

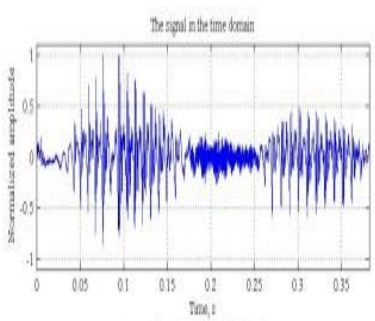


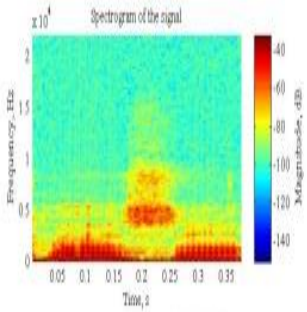


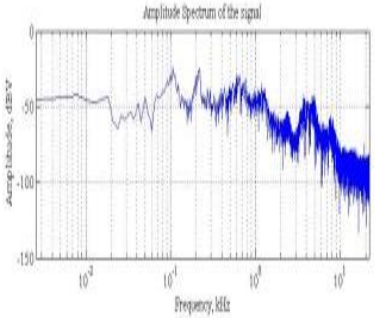
Results Audio: Ca

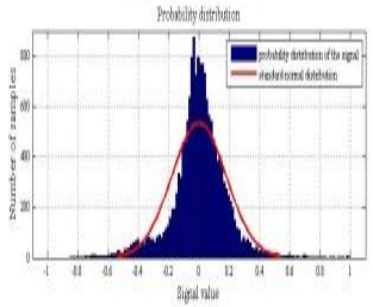
- Silabes

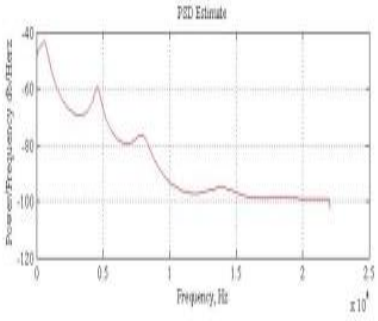
Results Audio: CASA

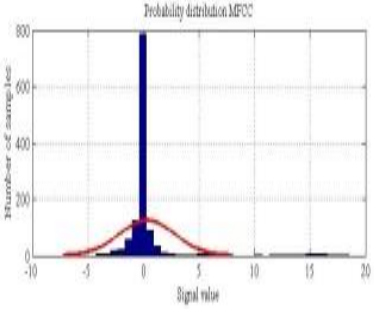


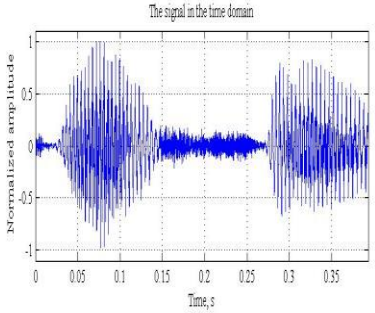


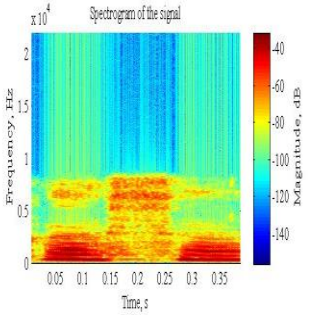


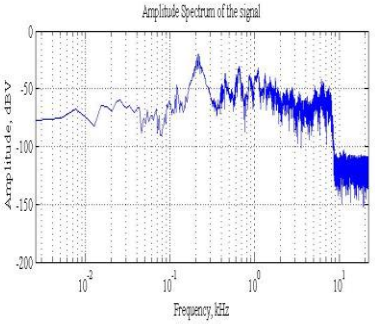


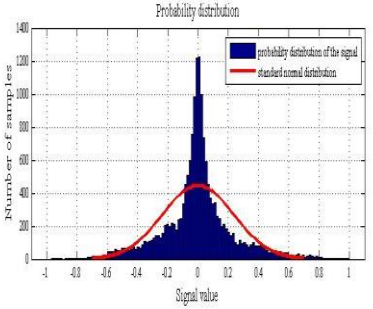


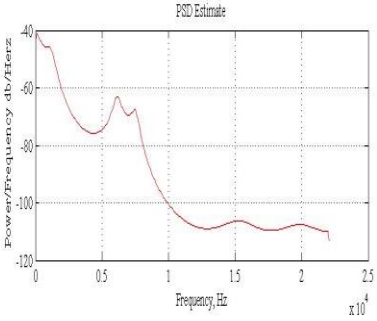


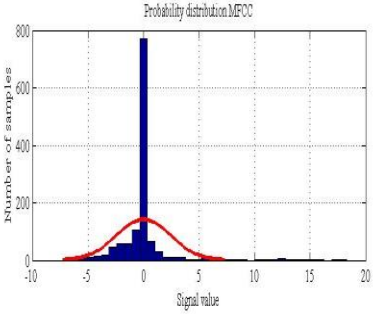










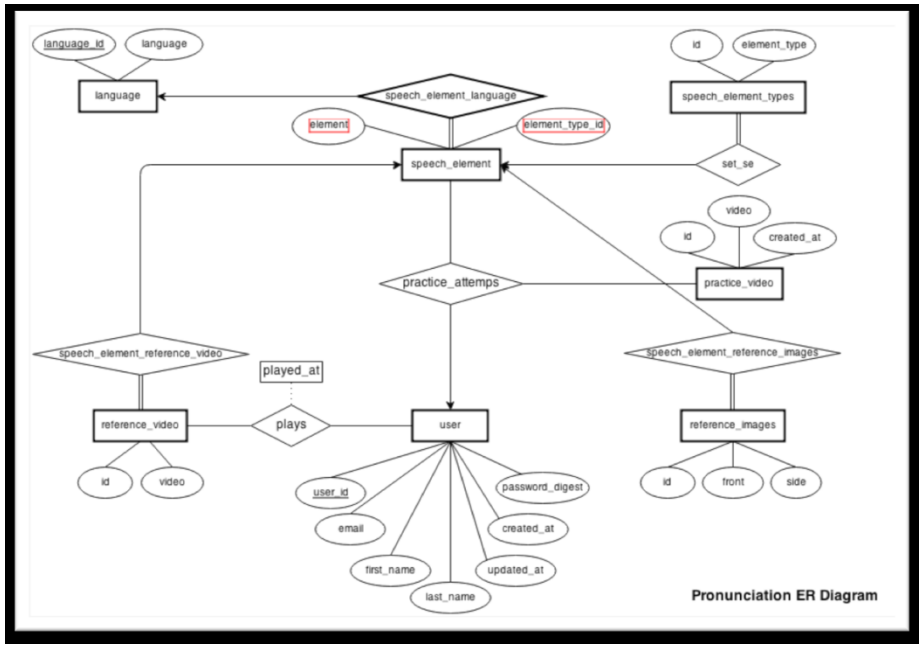


Results Audio: Word Segmentation into phones

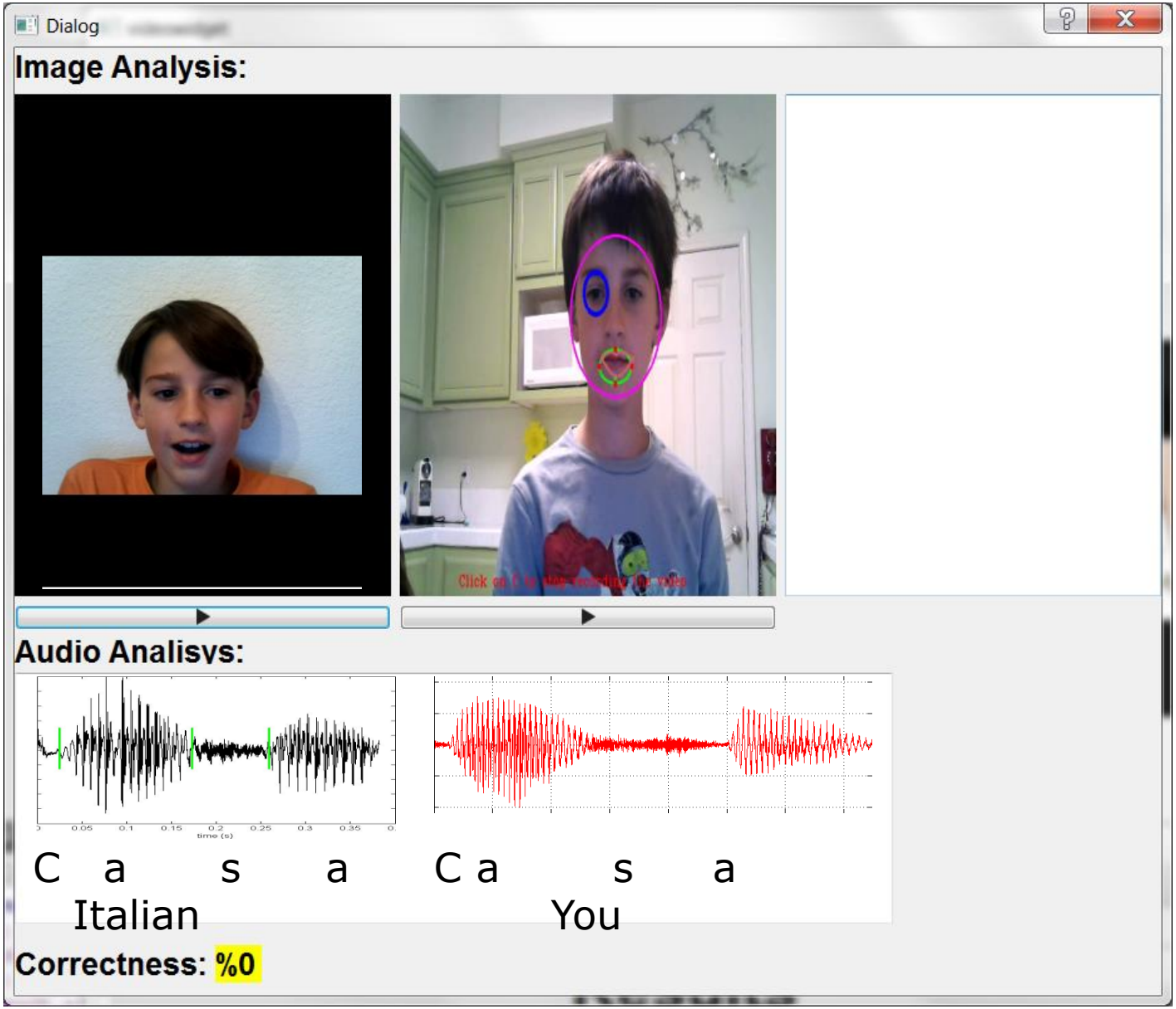
DataBase

- Around 5000 video/audio with “perfect” italian pronuntiation stored (right now only 100 hundred).
  - Using MariaDB Galera cluster in Amazon Web Services (AWS) by now
- Need to also store also videos from user , to replay. But this can be stored on user device
- Search and retrieval of the video is fast enough doesn’t seam too interesting to accelerate search on the GPU

Database Schema



Results


Dialog
Image Analysis:
Audio Analisys:
C a s a
Italian
C a s a
You
Correctness: %0

Conclusion

- Database currently stores rules for each word
- Should be change to automatically select rules that apply
- Combining Image and Audio help the pronuntiation
- The more you gesticulate the better we will give feedback
- Automatic evaluation and grading for educational pourposes

Questions