Spoken Digit Recognition

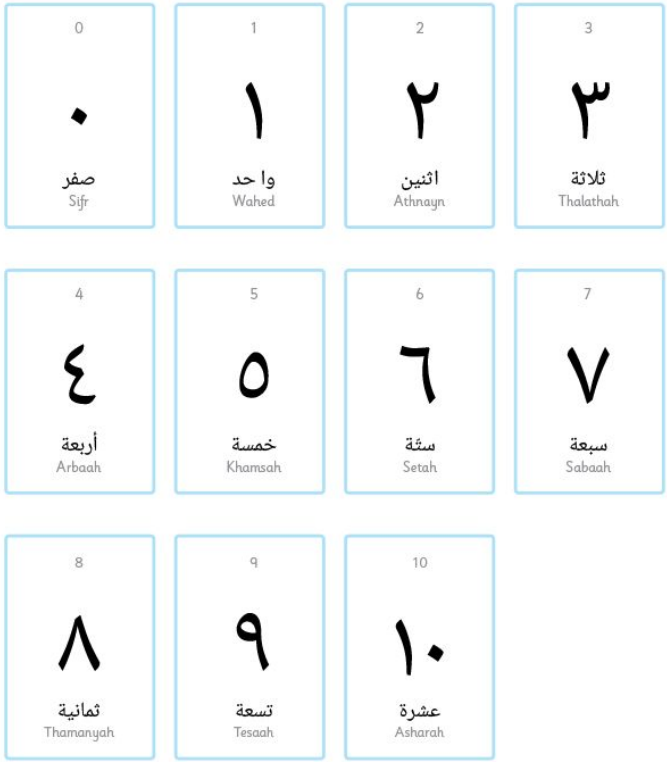
Ian Hanus

Background

- Advancement in speech recognition largely due to advances in AI/ML
- Mel-frequency cepstral coefficients (MFCCs) is one of the best known parametric representation of speech signals used in the speech recognition process
- MFCC is intended to artificially create the human ear's working system
- Systems utilizing MFCC predictors are currently used in identifying airline reservations, redirecting callers w/ spoken commands, and more
- There is no perfect speech recognition system: always room for improvement

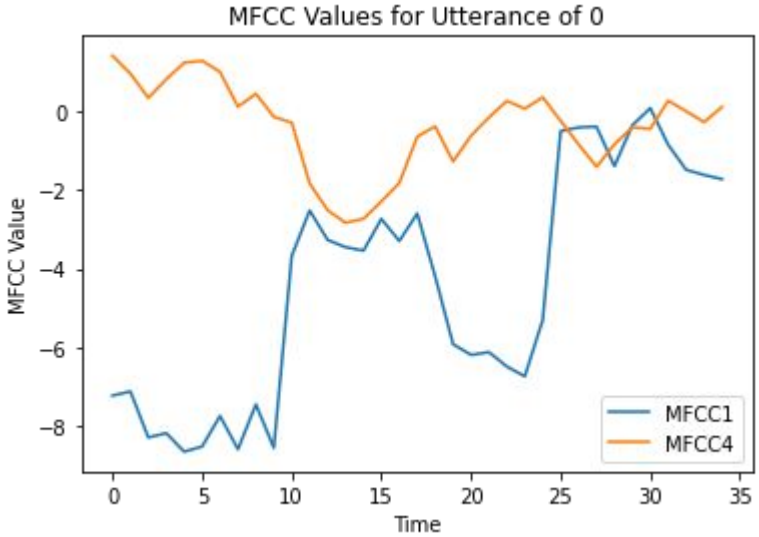
Project Goals

- Create an algorithm that correctly identifies spoken Arabic digits 0-9 with as few errors as possible, given the spoken digit's MFCC breakdown
- Compare the effectiveness of clustering by k-means to that of expectation-maximization using a metric of percent correct classification
- Examine the effect of speaker gender on predictor classification, and determine whether having separate models is necessary
- Determine the optimal number of components for characterizing digits in MFCC space
- Analyze and determine the effectiveness of certain MFCCs by including/excluding them in the classification process
- Especially interesting, as this is directly applicable to translation technologies. A user with no knowledge of Arabic could see what digit is being spoken, and this could possibly be extended to language translation



Data Exploration

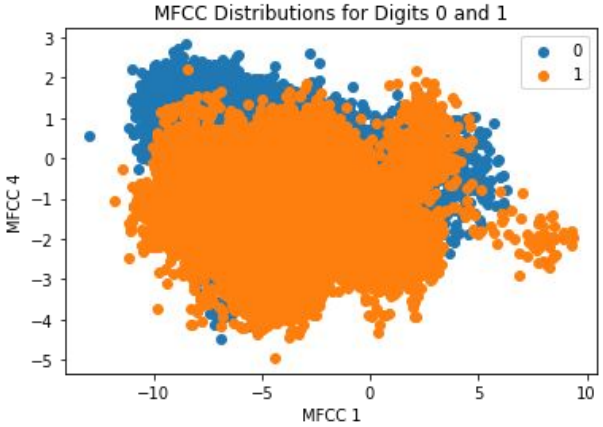
- Some MFCC values vary largely between phonemes, while some vary relatively little. The figure to the right shows the greater variance of MFCC 1 compared to that of MFCC 4 in the Arabic digit ‘sifir’
- Different digits have different numbers of phonemes
  - Distinct shifts in the cepstral coefficients correspond to phonemes in a digit
  - Could possibly correspond to an intuitive number of clusters in MFCC space
  - Number of phonemes were approximated below

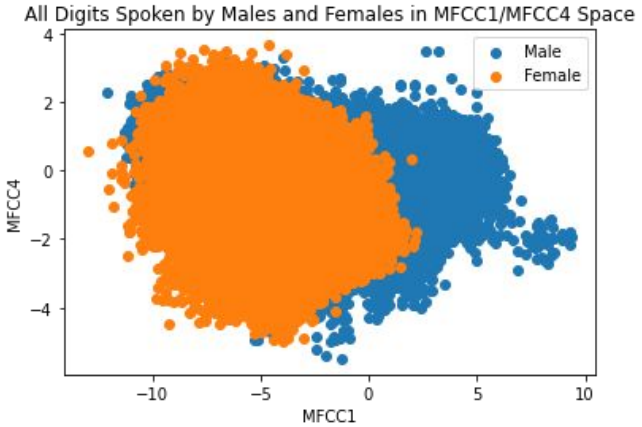


Digit	0	1	2	3	4	5	6	7	8	9
Approximate # of Phonemes	4	3	4	4	5	4	4	4	5	3

Data Exploration (cont.)

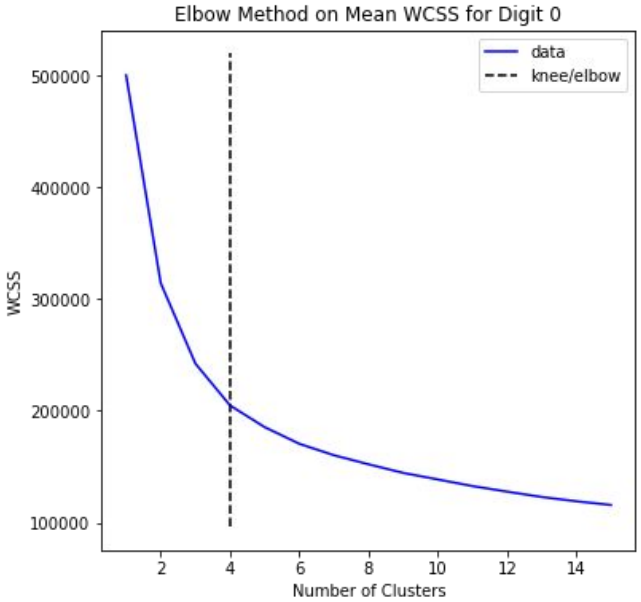
- Utterances have varying length in time, even within digits
  - Best to classify within solely ‘MFCC Space’
  - Although there is much overlap, the upper right figure shows a promising amount of variance between digits. Considering up to all 13 MFCCs, that variance between digits should be even more distinguished, allowing for better classification
  - Omission of time data does lead to less information, and could lead to confusion of words with similar phonemes in different orders
- Male and female voices tend to pronounce certain phonemes differently
  - This leads to a separation between the same digit spoken by male and female speakers
  - The effect of classification of speaker as male or female should be compared to use of the total data





K-Means Clustering

- K-means clustering is an unsupervised learning technique used to split data into a specified k clusters
- The optimal number of clusters can be found using the ‘Elbow Method’, which finds the point of diminishing returns on minimizing WCSS (Within-Cluster-Sum-Of-Squares) from increasing k
    - Visualization performed using PyPi kneed
- Originally, it was expected that the optimal k-value would be equivalent to the visible number of phonemes in a digit, but comparing the table below to the approximate number of phonemes shows only a slight correlation between the two
    - Digit 9 has only 3 phonemes but 5 optimal clusters



Digit	0	1	2	3	4	5	6	7	8	9
Optimal Component #	4	4	4	4	4	4	4	4	5	5

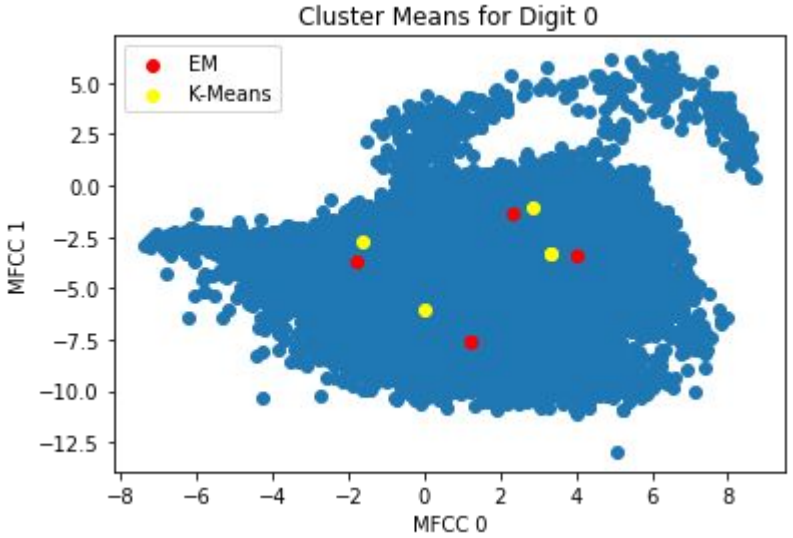
Classifying Using K-Means

- K-Means clustering was performed on each digit, recording the mean, weight (number of points in cluster/total number of points), and covariance for each of the clusters
  - sklearn KMeans was used to form clusters
- Classification was done first by maximum likelihood: likelihood of a test utterance being a certain digit, using the weights, means, and covariances calculated from the k-means clusters for those digits
  - scipy multivariate_normal was used to get $p(x_n|\Delta_{m,d})$
- Log-likelihood ended up being used to avoid numerical underflow, but as log is a monotonic function the digit corresponding to the maximum likelihood would also correspond to the log likelihood
- The digit corresponding to the maximum of these 10 log-likelihoods was taken to be the digit classification

$$\log(p(X|\Delta_d,\pi_d)) = \sum_{n=1}^{N} \log(\sum_{m=1}^{M} \pi_{m,d}\, p(x_n|\Delta_{m,d}))$$

Classifying Using EM

- Expectation-maximization models clusters using statistical distributions (in this case multivariate normal distributions), instead of the centroid model used in k-means
- Same maximum likelihood process applied, using the cluster weights, means, and covariances extracted from the EM clusters
    - sklearn mixture was used for expectation-maximization
- The plot to the right shows cluster centers calculated by each method: they may not seem entirely representative of clusters in the data, but this is because these particular cluster means are calculated across all 13 MFCCs



Testing Optimal Number of Components

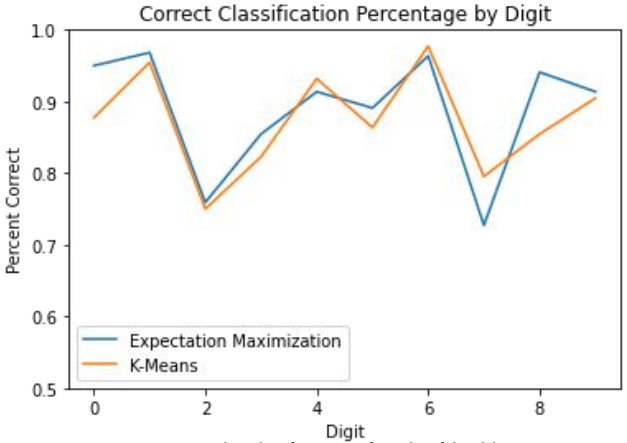
- Three different sets of number of components were tested using the k-means and EM classifiers
  - The median optimal number of components across all sets
  - The optimal k for each set
  - The estimated number of phonemes
- Optimal k for each digit chosen using the elbow method had the highest performance
  - Will be used throughout the rest of the experiment

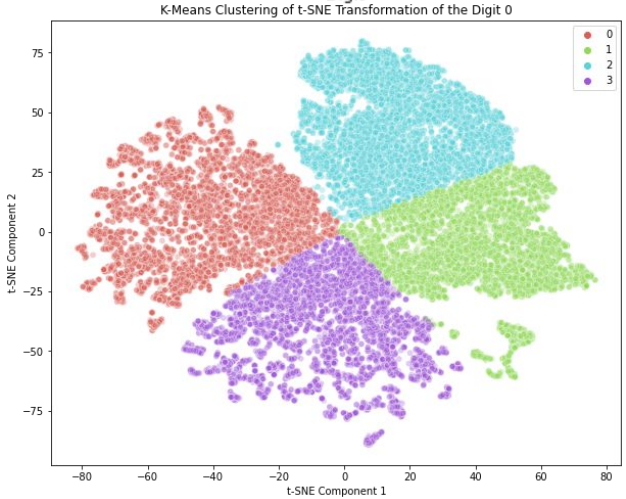
	0	1	2	3	4	5	6	7	8	9	Percent Accuracy
Median	4	4	4	4	4	4	4	4	4	4	0.864
Optimal K	4	4	4	4	4	4	4	4	5	5	0.873
# Phonemes	4	3	4	4	5	4	4	4	5	3	0.848

Comparison of Classifiers

Classification Method	Percent Correct
K-Means	0.873
Expectation-Maximization	0.888

- Results for the k-means and EM classification described in the previous slides are shown, using the ‘default’ parameters of all 13 MFCCs and optimal component number for each digit
- The two methods performed similarly, with the expectation-maximization slightly outperforming the k-means
- t-SNE confirms qualitative validity of clusters by reducing dimensionality from 13 to 2
- Note: digits 2 and 7 are the most misclassified
    - Could certain MFCCs or the latent male/female feature be causing this confusion?

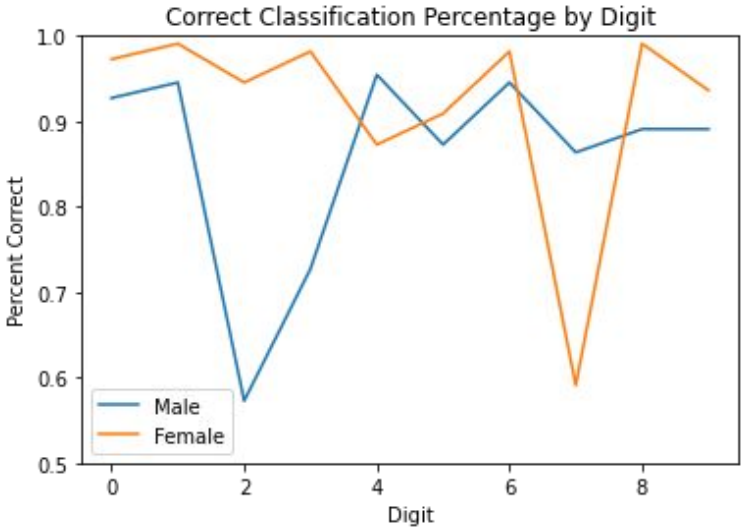




Male vs. Female

Binary Gender	Percent Correct
Male	0.859
Female	0.917

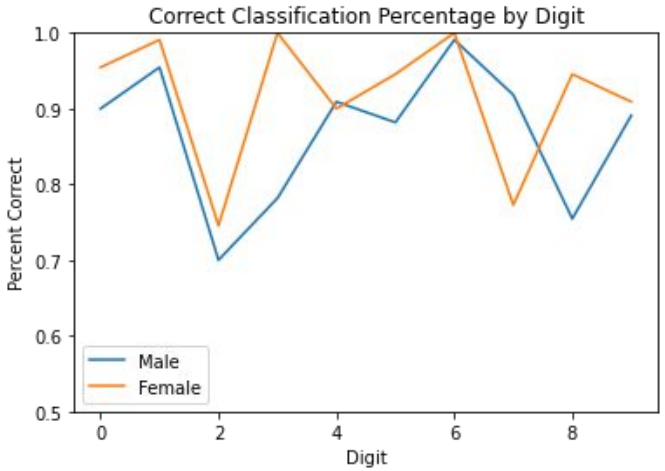
- The EM classifier trained across all data was evaluated separately on exclusively male and exclusively female test data
- The classifier performs better for female speakers than male speakers on the data
    - Interestingly in opposition to many modern speech recognition algorithms, which tend to have a higher error rate for women (Tatman 2017)
- The decreases in accuracy for digits 2 and 7 can still be seen here, but digit 2 only decreases for male-identifying speakers and digit 7 only decreases for female-identifying speakers



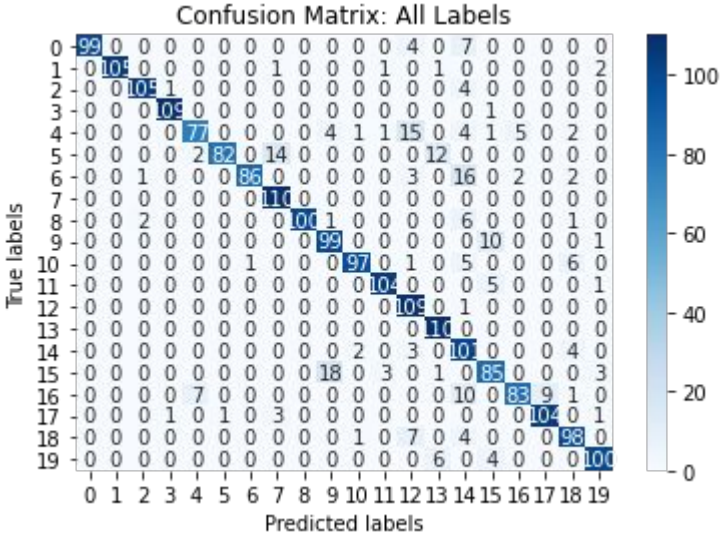
2 Approaches to Gender Consideration

- Altering GMM to have 20 likelihoods
  - True label of 0 is a 1 for males, true label of 1 is a 1 for females, etc.
  - Maximum likelihood estimate remained $\log(p(X|\Delta_d, \pi_d)) = \sum_{n=1}^{N} \log(\sum_{m=1}^{M} \pi_{m,d} p(x_n|\Delta_{m,d}))$
  - Changed to iterate over all 20 models, taking maximum likelihood, and mapping back to the original digit
- Train 2 separate classifiers, using exclusively male data on the first and exclusively female data on the second
  - Simple to choose which model if speaker gender is known
  - If not, a binary classifier can be implemented to first guess the gender of the speaker

Combined Gender Model



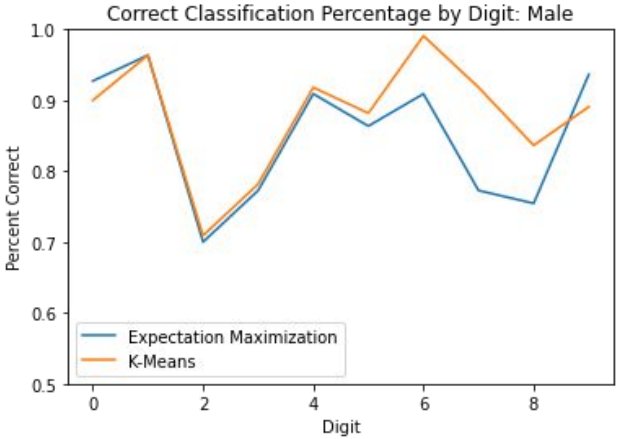
Data	% Accuracy
Male	0.878
Female	0.906
Total	0.892

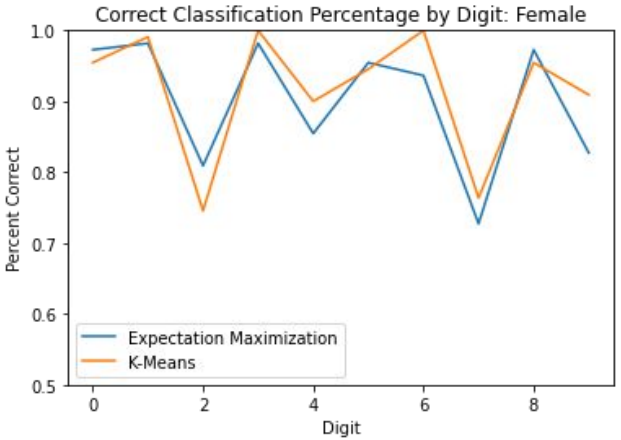


- Slight overall improvement in accuracy using the single MLE model across 20 likelihoods
- Male 2 and female 7 are misclassified significantly less often
  - Higher misclassification rates for female 2 and male 8
  - Very little of the misclassification is across gender (generally classified as different digit within same gender)
- Interestingly, model training time increase sixfold

Separate Gender Models

- Separate models were trained for the gender
- Overall higher performance when trained on data of exclusively one gender
    - Now, k-means outperforms expectation maximization
    - Combined percent accuracy of 0.907
    - If gender is known, more accurate to use gender-exclusive training: if not, dependent on how accurate a gender classification system would be. The minimum threshold of percent accuracy for such a gender classifier would have to be 0.970 (0.9776 achieved using MFCC and GMM in Yücesoy 2013)
    - Still slight dips at 2 for males and 7 for females





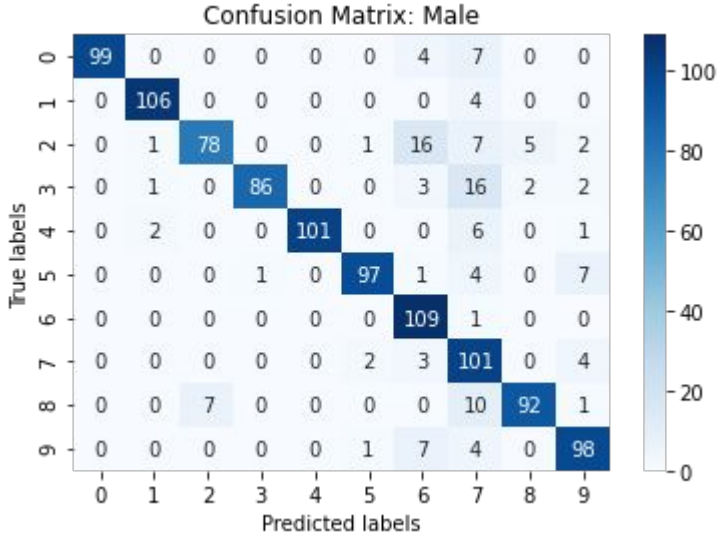
Male

Classifier	Percent Correct
EM	0.851
K-Means	0.897

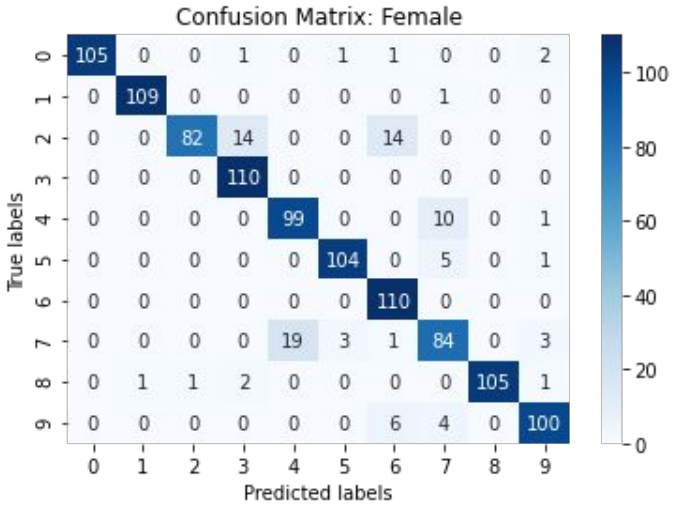
Female

Classifier	Percent Correct
EM	0.902
K-Means	0.916

Separate Gender Models (cont.)



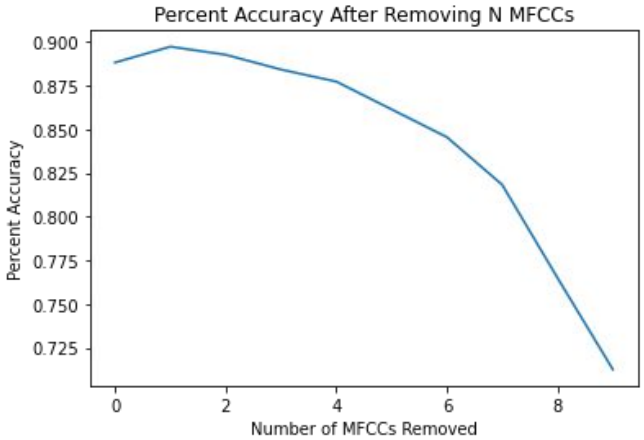
- Common mischaracterizations for males
  - 2s are labeled 6 or 7
  - 3s are labeled as 7



- Common mischaracterizations for females
  - 7s are labeled as 4s
  - 2s are labeled as 3s or 6s

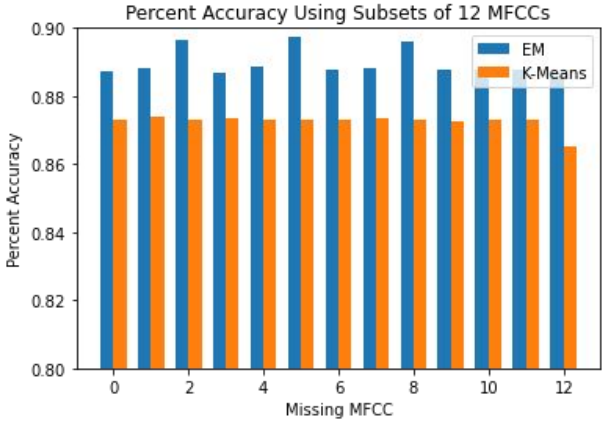
MFCC Subsets

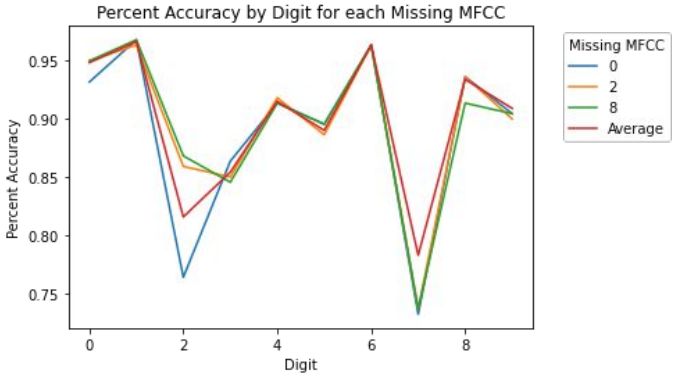
- The following was done using all data, without separation by gender
- Subsets of all possible MFCCs were examined using the following process
    - Starting with all 13 MFCCs, all 13 possible subsets of 12 MFCCs were created
    - K-means and EM classification were performed, and the subset w/ the highest percent accuracy was kept
    - All 12 subsets of 11 MFCCs from the kept subset of 12 MFCCs were created, and K-means and EM classification were performed again
    - Process was repeated until 9 MFCCs were removed, at which point there was an obvious drop in percent accuracy of the classification
- Maximum accuracy at excluding 1 MFCC



MFCC Subsets (cont.)

- The plot on the left shows that the maximum accuracy of 0.897 was obtained by removing MFCC from the training and testing data
- The largest dips in accuracy came from removing MFCC 0, MFCC 2, and MFCC 8
    - Removing MFCC0 leads to a large increase in misclassification of the digit 2
    - Removing any of the three leads to a large increase in misclassification of digit
    - Average loss over all MFCC removals was plotted for comparison
- Subset of MFCCs [0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12] minimizes classification error
- No removal of any digit resulted in higher prediction accuracy of digits 2 or 7





Gender Exclusive Model: MCCF Subset

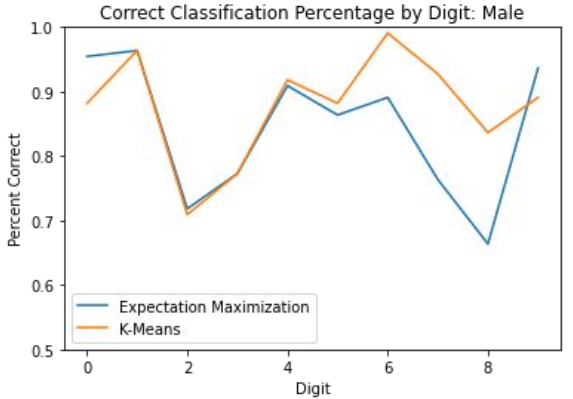
- The same subset exclusion was done on the separated male and female models
  - Least effective MFCC for males was MFCC 2
  - Least effective MFCC for females was MFCC5
- However, both gender-exclusive models had their performance decrease with the exclusion of an MFCC
- Gender-exclusive models still perform better with all 13 MFCCs than the original model does with the subset of 12 MFCCs

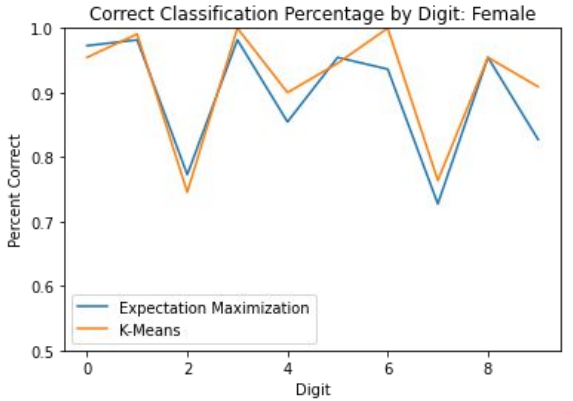
Male: Excluding MFCC2

Classifier	Percent Correct
EM	0.844
K-Means	0.877

Female: Excluding MFCC5

Classifier	Percent Correct
EM	0.896
K-Means	0.915





Conclusions

- The overall percent accuracy of the classifiers was maximized at 0.907, assuming speaker gender is known
  - Took speaker gender into consideration
  - Used all 13 MFCCs
  - Used optimal component numbers

Digit	0	1	2	3	4	5	6	7	8	9
# of Components	4	4	4	4	4	4	4	4	5	5

- Overall, this suggested model performs fairly well. One major caveat is how it performs on certain digits: the high misclassification rate on male speakers saying the digit 2 and female speakers saying the digit 7 means that this system should not be trusted with important tasks such as password entry or data validation
- The choice to use all 13 MFCCs was important, as it gave the classifiers as much information as possible to make a decision. Even though percent accuracy did increase on the overall data with the exclusion of one MFCC (possibly providing confounding data), the decrease in performance with the exclusion of more than one MFCC and sharp decrease after excluding three MFCCs shows how important the information they provide are
- The choice to have separate models for speaker genders was fairly important, allowing for an almost 3% increase in accuracy. This could be due to a difference in the way genders pronounce certain phonemes, or the differing average frequencies of the gender. There wasn't a huge difference between the 20 GMMs and the split gender models, so if speaker gender is not known it is likely best to use the combined gender model.

Conclusions (cont.)

- The modeling choice of number of components was also somewhat important. Although it made intuitive sense to use the number of phonemes as the number of components, by using the numerically optimal number of components percent accuracy was increased by 2.5%.
- One thing that I am a bit worried about is my decision-making process: it was almost entirely driven by percent accuracy on the testing data. If the testing data is not entirely representative of new data (i.e. a new speaker), then I am not confident that the classifier will perform well. It is possible that using the intuitive number of phonemes as number of components, instead of the computed optimal number of components, could be more useful with a wider range of test data.
- Lessons Learned:
    - Visualize as much as possible to have a better intuition as to what is happening
    - Examine tradeoff between increasing metrics and understanding the heart of a problem
    - Format data in a more intuitive way ASAP: lots of troubleshooting at the beginning trying to index through too disorganized of an array

Collaboration

- Code review of maximum likelihood estimation w/ Justin Kim
- Discussion of t-SNE, optimal k values, and speaker gender effects w/ Justin Kim
- Results comparison w/ Justin Kim
- Shared code for data-reading w/ Steven Cheng
- General project discussion w/ Cannon Palms, Hyunjae Lee, Justin Kim, Steven Cheng, and Will Carlson

References

Dave, Namrata. "Feature extraction methods LPC, PLP and MFCC in speech recognition." *International journal for advance research in engineering and technology* 1.6 (2013): 1-4.

Oppenheim, Alan V., and Ronald W. Schafer. "From frequency to quefrency: A history of the cepstrum." *IEEE signal processing Magazine* 21.5 (2004): 95-106.

Tatman, Rachael. "Gender and dialect bias in YouTube's automatic captions." *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 2017.

Yücesoy, Ergün, and Vasif V. Nabiyev. "Gender identification of a speaker using MFCC and GMM." *2013 8th International Conference on Electrical and Electronics Engineering (ELECO)*. IEEE, 2013.