# Applied Probability for Statistical Learning

**ECE 480**
**Fall 2021**

## Course Project: Recognizing Spoken Digits

**Although we all have unique voices and ways of saying words, there ie enough commonality in speech that we can almost always recognize speech independent of the speaker. The mechanisms by which human listeners recognize speech are not yet fully understood, though it is surely more complex than recognizing sequences of phonemes, the perceptually unique units of sound in a language. For example, "recognize speech" and "wreck a nice beach" are comprised of nearly identical phonemic sequences, yet most human listeners would be able to readily distinguish between the two phrases. Automated speech recognition systems, however, may not find distinguishing between these two phrases to be a trivial task. In this project, you will explore feature modeling for automated recognition of the spoken digits 0 through 9, spoken in Arabic.**

## Project Background

The focus of our course project is identifying which of the ten digits 0 through 9 was spoken, based on pre-computed cepstral cofficients. We are starting with cepstral cofficients so we can focus on modeling the features, rather than the signal processing required to generate the features.[1]

Each of the 10 spoken digits is comprised of a unique set of phonemes, so we expect each digit to be represented by a unique set of cepstral coefficient clusters. The phonetic pronunciations of the numerals in Arabic are:[2]

| | | | | |
|---|---|---|---|---|
| 0: sifir | 2: ithnayn | 4: araba'a | 6: sittah | 8: thamanieh |
| 1: wahad | 3: thalatha | 5: khamsa | 7: seb'a | 9: tis'ah |

These phonetic pronunciations provide domain knowledge that allow you to estimate the number of unique phonemes (or number of cepstral coefficient clusters) you anticipate for each digit.

The goal of this project is to explore a variety of probabilistic models for the cepstral coefficients and investigate the impact of modeling choices on subsequent maximum likelihood classification of the spoken digits.

[1] If the process of modeling speech via cepstral coefficients fascinates you, you can learn more about audio signal processing in ECE 485.

[2] Phonetic pronunciations provided by a former student, Dima Fayyad (ECE '20).

## Models of Cepstral Coefficient Distributions

For this project, we are going to model the distributions of the cepstral coefficients using Gaussian mixture models (GMMs). The component parameters for the GMMs can be found in a number of ways. We will discuss two this semester:

1. identifying clusters via K-Means and then calculating the mixture component parameters for each of the clusters, and
2. modeling the data as coming from a mixture model and using the expectation-maximization (EM) algorithm to estimate the mixture component parameters.

You should explore in your project the impact of both approaches to developing the GMMs on spoken digit recognition performance.

## Maximum Likelihood Classification

For this project we are going to restrict ourselves to maximum likelihood classification of the spoken digits. Exploring the choice of classifier is explicitly not a goal of this project; the goal of this project is to explore the effects of *modeling* choices.

The likelihood of a time series of $N$ frames of cepstral coefficients $\mathbf{X}$ given the $M$-component mixture model parameters $\mathbf{\Delta}_d$ and $\mathbf{\Pi}_d$ for the $d^{th}$ spoken digit is

$$p(\mathbf{X}|\mathbf{\Delta}_d, \mathbf{\Pi}_d) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_{m,d} p(\mathbf{x}_n|\Delta_{m,d}),$$

where each $\Delta_{m,d} = \{\mu_{m,d}, \Sigma_{m,d}\}$ represents the mean and covariance of the $m^{th}$ Gaussian mixture component for the $d^{th}$ spoken digit. The digit with the largest likelihood of the data given the digit's model is selected as the classification result.[3]

## Potential Modeling Questions

In addition to the approach for finding the mixture component parameters, there are other modeling questions that could be explored.[4]

- Which cepstral coefficients should be used in the model? All of them? Some subset?
- Should you constrain the GMM mixture component covariance estimates. If so, how should you constrain them? Assume diagonal covariances (independence among cepstral coefficients)? The same covariance for all cepstral coefficients? The same variance for all cepstral coefficients?[5]
- What is the impact of choosing how the frames are aggregated? What if $\mathbf{x}_n$ is a single frame? What if it is all frames concatenated? What if it is a subset of frames, such as 5 frames, or 1/4 of the total number of frames?
- How should the situation of different numbers of frames for each token be addressed?
- Should the latent variable 'speaker gender' be incorporated into the model?

[3] In the unlikely case of a tie among two or more digits, a classification result is randomly selected from the set of digits with the highest likelihood.

[4] This list is intended to jump-start your thinking about modeling questions that could be asked; it is not intended to enumerate all the modeling questions you should address in your project, nor is it an exhaustive list of all modeling questions that could be asked.

[5] Constraining the covariance matrices is reducing model flexibility, and so is moving the model along the bias-variance trade-off from a more flexible model, with greater variance, toward a more rigid model, with greater bias. Reducing model flexibility by introducing constraints on the model may also serve to improve the model parameter estimates, particularly in the case of insufficient data.

## Project Data

The dataset for our course project this semester is the Spoken Arabic Digit dataset, which is available from the UCI Machine Learning Repository.[6] This dataset contains a time series of 13 cepstral coefficients for 8800 unique speech tokens, where each token is a single utterance of one of the digits "zero" through "nine" (in arabic!). Each of the 10 digits is recorded 10 times by each of 88 unique speakers (44 female speakers and 44 male speakers). The 13 cepstral coefficients for each token are computed for a series of frames, thus producing a time series of cepstral coefficients for each token, with each token generally represented by 35-40 frames.

## Project Resources

You are not expected to write your own code to perform k-means clustering, expectation-maximization, or maximum likelihood classification. You may use toolboxes or packages that are available for Matlab or Python. Be sure to cite any packages or toolboxes you leverage.

## Collaboration

Even though you are each individually responsible for completing your own project and submitting your own Slidedoc describing your project, I strongly encourage you to collaborate extensively with others in the class. You may interact with your classmates in much the same way you would interact with a team: share and debate ideas, collaborate on code and share your code, and compare and contrast results and interpretations of results, as a few examples.

Every student is responsible for completing their own project and submitting their own Slidedoc describing their project efforts and results. There are two motivations for requiring individual submissions: 1) it is to your benefit to understand every aspect of the project, and 2) it is to your benefit to be able to continue making progress toward completing the project even if another student's personal circumstances limit their ability to engage with the project for a time.

[6] `https://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit`

## Recommended Project Milestones

*Week 1: (ends Friday 10/22)*

- Ensure you can load/read the data
- Ensure you can plot the data (cepstral coefficient as a function of window index), and you can see distinct shifts in the cepstral coefficients corresponding to transitions between phonemes in a digit
- Start slidedoc

*Week 2: (ends Friday 10/29)*

- Gaussian mixture model (GMM) estimation via k-Means
- Get started on maximum likelihood (ML) classification from k-Means GMM model
- Update and continue slidedoc

*Week 3: (ends Friday 11/5)*

- Continue ML classification from k-Means GMM model
- Explore additional modeling options
- Update and continue slidedoc

*Week 4: (ends Friday 11/12)*

- Gaussian mixture model (GMM) estimation via expectation-maximization (EM)
- Get started on maximum likelihood (ML) classification from EM GMM model
- Update and continue slidedoc

*Week 5: (ends Friday 11/19)*

- Continue ML classification from EM GMM model
- Explore additional modeling options
- Update and continue slidedoc

*Week 6: (ends Friday 12/3)*

- Finalize GMM estimation by k-Means and ML classification
- Finalize GMM estimation by EM and ML classification
- Explore additional modeling options
- Update and finalize slidedoc

*Monday 12/13* ***10:00PM****: Slidedoc due @ conclusion of our Final Exam time block*

- Note **submission deadline of 10:00PM** (not 11:59PM as earlier in the semester)
- Late submissions will not be accepted; not submitting the project report by 10:00PM 12/13 is equivalent to being absent from the final exam and will result in a grade of X.

## Report Slidedoc Guidance

A Slidedoc is a document that is more complete than a slide presentation, yet more concise than a conventional report. My goal in going to this format for the project documentation is to streamline the "reporting back" process – there is only one document for you to prepare instead of two (a report and a separate presentation), and your effort for this single document is focused on efficiently communicating the salient points. A Slidedoc that describes Slidedocs is available at: `https://www.duarte.com/slidedocs/`

There are no page requirements or limits for the Slidedoc; it should be parsimonious – as long as it needs to be to fully describe what you have done, but no longer than necessary.

Each component of the Slidedoc described below should be complete and thorough, so that someone reading your Slidedoc should have a good understanding of what you are describing solely from your descriptions of it. References to relevant outside sources you used to support your project should be provided, and citations must be provided for any ideas, thoughts, statements, pictures, or figures that are not your own.

***If a component is missing from the Slidedoc, then the corresponding score for that component is necessarily "N" (i.e., 0 points).***

### *10% Clarity and Organization*

This is a formal document. As such, writing (organization, sentence structure, etc.) matters, and the presentation clarity and organization score will reflect the quality of the written presentation. I will not be specifically reading for grammar, spelling, etc., but I will notice if my comprehension is impeded by these elements and the Clarity and Organization score will reflect this. You are free to make use of the writing studio to help you improve your report.
`http://twp.duke.edu/twp-writing-studio`

**Why does clarity and organization matter when this isn't a writing class and I am not a writing teacher?** One of my jobs is to help you prepare for your professional career. Communication is a large component of many professional roles, so it is to your advantage to make the most of opportunities such as this to continue strengthening your communication skills. In many professional settings, strong communication skills are necessary for professional advancement, such as being appointed project lead or earning a promotion.

Present a Slidedoc that is clearly written, easy to follow, and complete as a stand-alone document, as would a Slidedoc delivered to a customer who hired you to complete this project.[7] Someone who is not taking (or has not taken) this class, but is familiar with the mathematical background, should be able to read your report and understand conceptually what you have done. Your presentation should describe the problem you are solving, describe your methods/approach to solving it, present your results, and present conclusions based on your results. This information should be presented in a logically organized, and sequential, way. Concepts should be defined or explained before they are used, and each page (slide) should have a key take-away point.

[7] This project completed as part of our class this semester may become an example you share during a future interview!

**How are Slidedocs different from slide decks?** Slidedocs are intended to be read, whereas slide decks (presentations) are intended to be heard.

While a well-designed presentation (slide deck) is typically highly visual with very few words (often organized as bullet points), a well-designed Slidedoc includes prose (full sentences organized into short paragraphs). While the prose in a Slidedoc may be more concise (and scannable) than the prose in a long-form report, it is still prose, not bullet points. This means the reader should be able to fully understand the message the page conveys solely from reading the words on the page; the reader should not need to imagine additional dialogue (as would be provided by a presentation speaker) to fully understand the message the page conveys.

You can think about what the "speaker script" for a presentation slide might be, and include that script as prose on the Slidedoc page. If you want to 'test' your page to see if it's a page from a Slidedoc or a page from a slide deck, read it out loud (only the words written on the page). Does it sound like a natural portion of a conversation? If it does, you have text for a good Slidedoc page. (Text alone does not necessarily make a good Slidedoc page; a good Slidedoc page typically also includes visual aid(s).) If, instead, the text on the page sounds like a series of disjoint statements, you do not (yet!) have text for a good Slidedoc page.

When I am reading the Slidedocs, I will read the words on the page; I will not imagine additional dialogue that may surround those written words if the Slidedoc were to be presented orally.

Specific pieces of advice:

1. Define acronyms the first time they are used.
2. Design slides titles to orient the reader to where they are in the Slidedoc.
3. Make use of spell-checkers and grammar-checkers.

### *10% Visualizations*

Support your textual content with visualizations. It may be helpful to provide visualizations that illustrate the efficacy of your model inversion (decay rate estimation) and your classification process.

Examples of plots you may choose to present include (this is *not* an exhaustive list of all the plots you could, or should, include):

- A few example model inversions, showing the measured time series and the predicted time series using the estimated model parameters.
- Scatter plots of the estimated decay rates, encoded by system type.
- Plot of the decision statistic surface for your classifier, with a scatter plot of the estimated decay rates, encoded by system type, superimposed on top.
- Cross-validated performance prediction (ROC curve) using the training data, to show how robust you expect your classifier to be when it is applied to the blind test data.

Your figures are expected to look professional. This means, at a minimum:

- Do not "Print Screen", screen capture, or snip/clip a figure window, as this approach results in low quality images (doing so will result in point deductions) . Instead, export/save the figure as a graphics file and then import the high quality image into your document.
- Label all axes.
- Include a legend.
- Include a descriptive title.

- Ensure all text is large enough to be readable after the figure is imported into your document. 8-point font is generally accepted as the smallest usable font. (Making the figure window smaller prior to exporting the figure generally results in larger fonts in the exported figure.)
- To reiterate: **Do not "Print Screen", screen capture, or snip/clip a figure window.** Instead, export/save the figure as a graphics file and then import the high quality image into your document.

The elements I am looking for are:

- Figures are not "print screen" images (that include the OS window frame and/or are pixelated)
- Axis labels
- Legends
- Descriptive titles
- Color/symbol/line type choices that facilitate disambiguating different curves or clusters

### *12% Problem Description*

Describe the goals of the project, and what is hoped to be achieved or gained at the completion of this project. If you assimilate contextual information from reading the background and introductory sections of other resources, then those resources are references for your problem description.

The elements I am looking for are:

- What are the project goals?
- Why is this an interesting or important problem?
- Why might others be interested in this problem?
- *Example of additional exploration/investigation:* Describe a scenario related to your engineering or technical interests (other than speech recognition which is the focus of this project) for which a similar modeling framework may apply.

### *14% Data (Feature) Modeling*

Provide descriptions of and motivations for modeling choices you explored, including representative visualizations that illustrate the motivations for and/or impacts of various modeling choices. Also include the mathematical representation (equation) for each of your modeling choices. Someone else should be able to replicate your modeling processes from your descriptions of them (without necessarily having access to the same toolboxes, packages, libraries, etc. that you may have used).

The elements I am looking for are:

- What modeling choices did you explore?
- Why did you explore these modeling choices?
- What are the advantages / disadvantages of the modeling options?
- Key equation(s) that describe(s) the data (feature) modeling.
- Visualizations that illustrate motivations for and/or impacts of modeling choices.
- *Example of additional exploration/investigation:* Explore other modeling choices not outlined in the section describing potential modeling questions (*e.g.* threshold the covariance so that if a covariance is less than a threshold then the cepstral coefficients are assumed to be independent, should each speaker have a model, or speakers be grouped so cohorts of speakers each have a model).

### *14% Maximum Likelihood Classification*

Provide a description of how maximum likelihood classification is implemented for spoken digit recognition under the specific modeling choices you selected for exploration. Be sure to include the mathematical representation (equation) for your maximum likelihood classifier, as well as any variations of it that may arise under different modeling choices. Someone else should be able to replicate your classifiers from your descriptions of them (without necessarily having access to the same toolboxes, packages, libraries, etc. that you may have used).

The elements I am looking for are:

- Description of maximum likelihood classification.
- Why is maximum likelihood classification well-suited for this problem?
- Equations that mathematically describe maximum likelihood classification.
- What challenges did you encounter when implementing and/or applying maximum likelihood classification, and how did you overcome them?

### *14% Classification Performance Results*

The Slidedoc is expected to describe and show the results of your efforts toward modeling and classifying based on the cepstral coefficient features. Provide a quantitative description of how well the system performs under various modeling choices (*i.e.*, include confusion matrices, probability of correct digit, etc.), and in the text of your report interpret the results. It is insufficient to merely present a series of figures. Instead, talk the reader through the figures to ensure the reader is guided toward interpreting the figures in the way you intend for them to be interpreted and observing the key take-away points in the figures.

The elements I am looking for are:

- Quantitative results are presented (Confusion matrices and probability of correct digit).
- Quantitative results are described.
- Quantitative results are interpreted.
  - For what sub-cases (particular digits) does the classifier perform well? Why is this the case?
  - For what sub-cases (particular digits) does the classifier perform poorly? Why is this the case?

### *15% Conclusions*

Provide your overall assessment of this modeling exploration and the spoken digit recognition system you built, including what you see as strengths and weaknesses of the modeling and classification approaches, and your subjective assessment of the quality of the digit recognition results you obtained.

The elements I am looking for are:

- What modeling choices are important, because they have a large impact on spoken digit classification performance? Why is it that these modeling choices significantly impact performance?
- What modeling choices are less important, because they do not significantly impact spoken digit classification performance? Why is it that these modeling choices do not significantly impact performance?
- If you had to specify a single system (model and maximum likelihood classifier), what would that system be?
- How well does this modeling/classification system do for spoken digit recognition?
  - What are great things about the system?
  - What would you do to improve the system?

- What “lessons learned" will you carry forward with you?
  - What would you do differently next time?
  - What worked really well, and you would do the same way next time?

### *5% References*

References/citations must be included. For example, hierarchical clustering and k-means clustering are well-established unsupervised clustering techniques, so your description of the mathematical formulation must include citations to indicate to the reader that you are not the originator of the clustering technique. The references/citations must be books or journal articles. Websites are not suitable references, nor are our class lecture notes.

If you reproduce an image from another source (including websites) you must provide a citation for that image.

You must provide a citation for every toolbox or package you leverage. (We need to know what your code sources are.)

The elements I am looking for are:

- Citations for background or contextual information.
- Citations for GMM model estimation (k-Means and expectation-maximization(EM)).
- Citations for maximum likelihood classification.
- Citations for visualizations taken from other sources (including websites).
- Citations for toolboxes or packages you use.

### *3% Collaborations*

Describe your collaborations with other students in the class while working on this project.

The elements I am looking for are:

- Who did you share and debate ideas with while working on this project?
- Who did you share code with while working on this project?
- Who did you compare results with while working on this project?
- Who did you help overcome an obstacle while working on this project?
- Who helped you overcome an obstacle while working on this project?

### *3% Weekly Project Progress Report Surveys (submitted in Sakai)*

A short (4 question) progress report survey is due by the conclusion of each week – 11:59PM Fridays. Each progress report survey will be made available at noon on the Thursday preceding its submission deadline.

This project progress report is intended to help me understand how you are progressing toward completing the project. Toward this end, your honest self-assessment of your progress to date will be most helpful for helping us help you.

The questions in the Sakai survey are:

- What is your progress toward this week’s milestones?
- What questions remain to be answered to support your continued progress toward completing this week’s and/or next week’s milestones?
- What is your plan for continuing to make progress between now and next week’s progress report?

- What is your best estimate of the % complete for each milestone?

This progress report is scored as complete (all questions answered) or incomplete (any questions not answered or survey not submitted).

There are six weekly progress reports for this project.