Applied Probability for Statistical Learning

ECE 480
Fall 2024

Course Project: Recognizing Spoken Digits

Although we all have unique voices and ways of saying words, there is 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 pre-computed 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. 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 for estimation of the number of unique phonemes that are anticipated for each digit. (Think about how this domain knowledge may inform your choice of the number of mixture model components 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 *Digital Audio and Acoustic Signal Processing* (ECE 485) and *Natural Language Processing* (ECE 684).

[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 investigate two approaches in this project:

1. identifying clusters via k-means and then assuming the k-means clusters provide a good approximation to the mixture components for a GMM (calculating the GMM mixture component parameters from the k-means clusters), and
2. modeling the data as generated by a Gaussian mixture model (GMM) and using the expectation-maximization (EM) algorithm to estimate the GMM 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 C-dimensional frames of cepstral coefficients, $\mathbf{X}$ (whose dimensionality is $[N \times C]$), given the M-component mixture model parameters $\mathbf{\Delta}_d$ and $\mathbf{\Pi}_d$ for the d^{th} spoken digit is

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

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 that maximizes the likelihood of the data given the digit's model is selected as the classification result (the maximum likelihood estimate of the digit).[3]

Potential Modeling Questions

In addition to the approach for finding the mixture component parameters (either estimated from k-means clusters or via the EM Algorithm), 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)? Assume the same covariance for all cepstral coefficients (a tied covariance)? Assume the same variance for all cepstral coefficients (a spherical covariance)?[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?

[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.

- Should the varying numbers of frames for each token be addressed?
- Should the latent variable 'speaker gender' be incorporated into the model?

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.

There are two data files: `Train_Arabic_Digit.txt` and `Test_Arabic_Digit.txt`. You will use the training dataset, `Train_Arabic_Digit.txt`, to estimate parameters for your classifier, and the testing dataset, `Test_Arabic_Digit.txt`, to evaluate classifier performance.

The final two paragraphs of the "Data Set Description" document describe how the data are stored in the text files. There are no data files containing target variables or meta-data (*e.g.*, digit, speaker gender); meta-data are embedded in the data organization.

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 your computing platform of choice, and long as you cite any packages or toolboxes you use.

Collaboration

You are each individually responsible for completing your own project and submitting your own document describing your project. Even so, 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 project document 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 Thursday 10/17) Checkpoint: **HW #6 (10/24)**

- Ensure you can load/read the data
- Ensure you can plot the data (cepstral coefficient as a function of frame index), and you can see distinct shifts in the cepstral coefficients corresponding to transitions between phonemes in a digit
- Ensure you can pair-wise scatter plot the data, and you can see clusters in pair-wise combinations of the cepstral coefficients corresponding to phonemes in a digit
- Start project document

Week 2: (ends Thursday 10/24) Checkpoint: **HW #7 (10/31)**

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

Week 3: (ends Thursday 10/31) Checkpoint: **HW #8 (11/7)**

- ML classification from k-means GMM model
- Explore additional modeling options
- Update and continue project document

Week 4: (ends Thursday 11/7) Checkpoint: **HW #9 (11/14)**

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

Week 5: (ends Thursday 11/14)

- ML classification from EM GMM model
- Explore additional modeling options
- Update and continue project document

Week 6: (ends Thursday 11/28)

- Finalize GMM estimation by k-means and ML classification
- Finalize GMM estimation by EM and ML classification
- Explore additional modeling options
- Update and finalize project document

Thursday 12/5 ***3:00PM****: Slidedoc due*

- Late submissions will be accepted through 3:00PM Thursday 12/12
  - Late submission penalty waived through 3:00PM Tuesday 12/10
  - Submission by 3:00PM Wednesday 12/11 = 1 point penalty (1 letter grade)
  - Submission by 3:00PM Thursday 12/12 = 2 point penalty (2 letter grades)
- The 6 hour grace period applies to all submission deadlines listed above. Not submitting the project project document by 9:00PM Thursday 12/12 (end of the grace period on final day late submissions are accepted) is equivalent to being absent from the final exam and will result in a grade of X.

Project Reporting Guidance

Your project report format may be either a slidedoc or an ePortfolio, both of which are described below. These reporting formats will lead to a document that is complete in its presentation of the project effort and results yet is more concise than a conventional long-form report, thereby supporting your professional development by serving as an easily digestible artifact of your experience that you can share with others. The document that you author has the potential to be a great resource for you in the future, as a leave-behind document for interviews as well as a memory-refresher document for yourself in preparation for interviews or if you encounter a similar machine learning problem in the future.

My goal in requiring one of these two formats for the project documentation (slidedoc or ePortfolio) is to streamline the "reporting back" process, by focusing your effort for this document on efficiently communicating the salient points and to align this effort with activities that support your professional development. There are no length requirements or limits for your project report; it should be parsimonious – as long as it needs to be to fully describe what you have done, but no longer than necessary.

A traditional long form report is not a suitable project report format.

Whether you opt for the slidedoc or the ePortfolio format, each component of the project reporting described below should be complete and thorough in your document, so that someone reading your document should have a good understanding of what you are describing solely from your written descriptions in your document.

If a component is missing from the document, then
the corresponding score for that component is necessarily 0.

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.

Project Report Option: Slidedoc

A slidedoc that describes slidedocs is available at: `https://www.duarte.com/slidedocs/`

A common question is, "How are Slidedocs different from slide decks?" The short answer is slidedocs are intended to be *read* (all the words on in the slidedoc), whereas slide decks (presentations) are intended to be *heard* (words are missing from a slide deck because they are spoken by the presenter).

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.

Your slidedoc should stand completely on its own,
without you there to orally present it to the audience.

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 a slidedoc, I will read the words on the page; I will not imagine additional dialogue that may surround those written words if the slidedoc were presented orally.

Project Report Option: ePortfolio

An ePortfolio is a digital presence in which you tell the story of your experiences, including reflection on what you learned and how the experiences shaped you.

The Integrative Learning Portfolio Lab at Stanford[7] maintains a wealth of information about ePortfolios, as well as a gallery of examples, on their website. Many of the ePortfolios shared through this gallery feature multiple experiences. If you choose to document your course project through an ePortfolio, it would become one experience, of hopefully many experiences, in your ePortfolio.

As for slidedoc reports, I will read the words on the page; I will not imagine additional narrative content that is not present on the page – ensure your narrative presentation is complete.

Academic Integrity Expectations

It is expected that your project document represents your own ideas and your own understanding of the concepts written in your own words. While the content is expected to be written in your own words, you may consult resources and references provided that 1) proper citations are provided for ideas you paraphrase, and 2) proper quotations are used and proper citations are provided for resources and references you quote.

It is *not* acceptable to:

- Copy content (text and/or images/figures) from class materials, including the project guidance, even if properly quoted and cited — do not copy from class materials. Material copied from class materials will be disregarded – treated as if it is not present in the document.
- Copy content (text and/or images/figures) from another resource/reference without both proper quotation and proper citation.
- Paraphrase content (text and/or images/figures) from another resource/reference without proper citation.
- Stitch together sequences of properly quoted and cited phrases and/or sentences to create passages that are not expressed in your own words.

Project Reporting Components

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 document: `http://twp.duke.edu/twp-writing-studio`.

[7] `https://eportfolio.stanford.edu/`

Why does clarity and organization matter when this isn't a writing class and I am not a writing teacher?

> An engineer who can't communicate works for one who can.
> — attributed to Neal Lerner, who was at MIT at the time and is now Chair of the English Department at Northeastern University.

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 document that is clearly written, easy to follow, and complete as a stand-alone document, as would a document delivered to a customer who hired you to complete this project. Someone who is not taking (or has not taken) this class, but is familiar with the mathematical background, should be able to read your document and understand what you have done without your oral presenting the document to them. Your document should describe the problem you are solving, describe your methods/approach to solving it, present your results, and present conclusions drawn from your results. This information should be presented in a logically organized, and sequential, way. Concepts and acronyms should be defined or explained before they are used, page titles (slidedoc) or section headings (ePortfolio) should orient the reader to where they are in the document, and key take-away points should be clearly articulated.

10% Visualizations

Support your narratives with visualizations. You should think about providing visualizations that illustrate the efficacy of your model 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):

- Time series plots of MFCCs as a function of frame index.
- Time series plots of MFCCs as a function of frame index, with estimated cluster ID for each frame encoded in the visualization.
- Pairwise scatter plots of the MFCCs.
- Pairwise scatter plots of the MFCCs, with the estimated cluster ID for each frame encoded in the visualization.
- Validated performance prediction (confusion matrices) – classifier trained on the training set and tested on the testing set.

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

5% 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.

30% Data (Feature) Modeling
(25% "Doing" the modeling; 5% "Reporting" on the 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, as well as 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).

Also include results illustrating the modeling outcomes of your models, such as visualizations that embed the clusters resulting from applying the models.

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 modeling choices.
- Visualizations that illustrate 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).

30% Maximum Likelihood Classification
(25% "Doing" the classification; 5% "Reporting" on the 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 document is also expected to describe and show the results of your efforts toward 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 document 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:

- 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?
- 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?

5% 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, k-means clustering and expectation-maximization are well-established modeling 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.

5% Collaborations

Describe your collaborations with others 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?

If you did not collaborate with others while working on this project, explicitly state that you worked entirely independently, and

1. explain why you chose to work independently, and
2. describe the resources you relied upon in the absence of collaborators.

Point Allocation and Scoring Criteria

"Doing" the ML and "Reporting" on the ML are considered separately for scoring, as shown in the following table.

	"Doing" the ML	**"Reporting" on the ML**
Clarity & Organization	–	10%
Visualizations	–	10%
Problem Description	–	5%
Data (Feature) Modeling	25%	5%
Maximum Likelihood Classification	25%	5%
Conclusions	–	5%
References	–	5%
Collaborations	–	5%
	50%	50%

“Doing” the ML is generating the results – writing the code and generating figures to display the results.

“Reporting” on the ML is explaining what was done to generate results by providing necessary background (*e.g.*, theory, equations) to explain how the results were generated as well as presenting and interpreting the results by providing narrative prose (text) that describes what the results represent and what the audience should observe in the results or learn from the results.

A pre-requisite for providing an interpretation of the results is generating results to interpret – a corresponding good faith effort toward “doing” the ML must accompany responses to explain/interpret the ML (“reporting” on the ML).

Scoring guidelines for the various elements are provided in the following table.

	“Doing” the ML	**“Reporting” on the ML**
100%	Exceptional/insightful setup/implementation	Exceptional insight/explanation **and** full and complete response
95%	Complete **and** correct setup/implementation	Full **and** complete response
90%	Minor shortcomings in setup/implementation	Minor shortcomings in response
85%	Shortcomings in setup/implementation	Shortcomings in response
80%	Significant shortcomings in setup/implementation	Significant shortcomings in response
75%	Major shortcomings in setup/implementation	Major shortcomings in response
65%	Severe shortcomings in setup/implementation	Severe shortcomings in response
50%	Catastrophic shortcomings in setup/implementation, with demonstrated good faith effort	Catastrophic shortcomings in response, with demonstrated good faith effort
30%	Catastrophic shortcomings in setup/implementation, without demonstrated good faith effort	Catastrophic shortcomings in response, without demonstrated good faith effort
0%	No implementation / No submission	No response / No submission