Speech Recording based Language Recognition

Leopold Cambier, Cindy Orozco Bohorquez, Matan Leibovich

lcambier@stanford.edu, orozcocc@stanford.edu, matanle@stanford.edu

Institute of Computational and Mathematical Engineering

Overview

- **Motivation**: Construct a real time language classifier for communication purposes.
- **Method**:
  - Construct ML estimator based on Gaussian Mixture Model (GMM) density estimations.
  - Feature vectors based on Shifted Delta Cepstral Coefficients (SDC)
- **Results**:
  - Classification error decreases with model complexity to a certain limit.
  - Optimal number of gaussians varies significantly across languages.
  - There is clustering pattern in the classification.

Data

Source: `https://community.topcoder.com/longcontest/?module=ViewProblemStatement&rd=16555&pm=13978`.

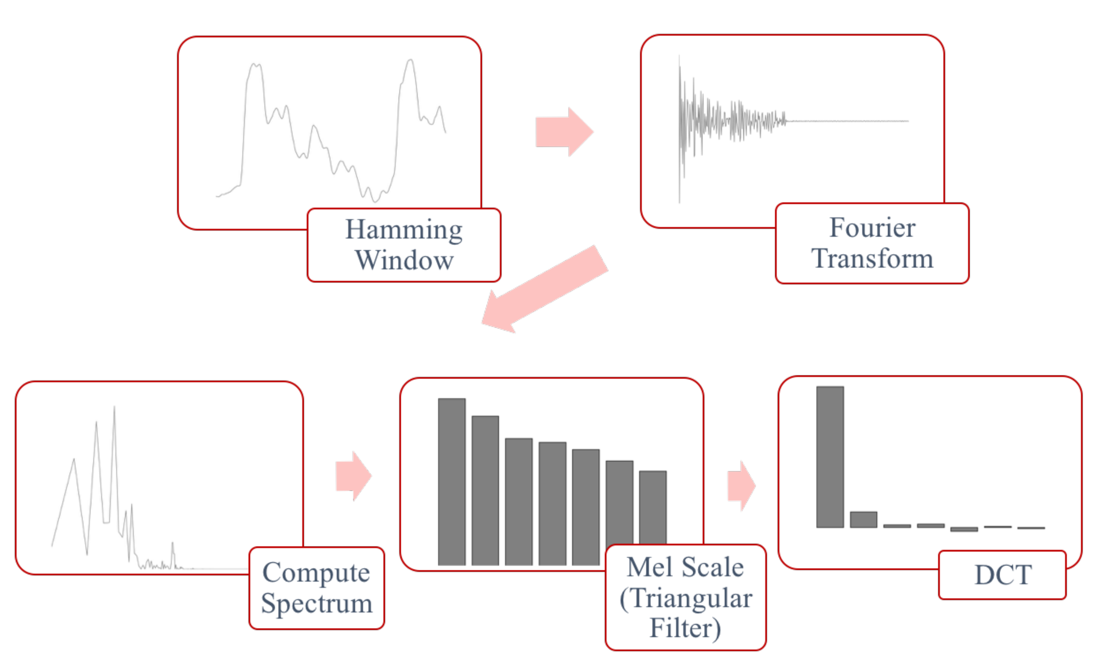
- $\mathcal{L}$: 176 different languages (some very exotic!)
- 376 10 sec. samples for each language
- Each sample divided into 210 ms long segments

Feature Extraction: Shifted Delta Cepstral (SDC)

❶ **Hamming Window**: $x_{\text{Ham}}(t) = (x^{*}\text{Ham})(t)$

❷ **Fourier Transform**: $\hat{x}(f) = \mathcal{F}(x_{\text{Ham}}(t))$

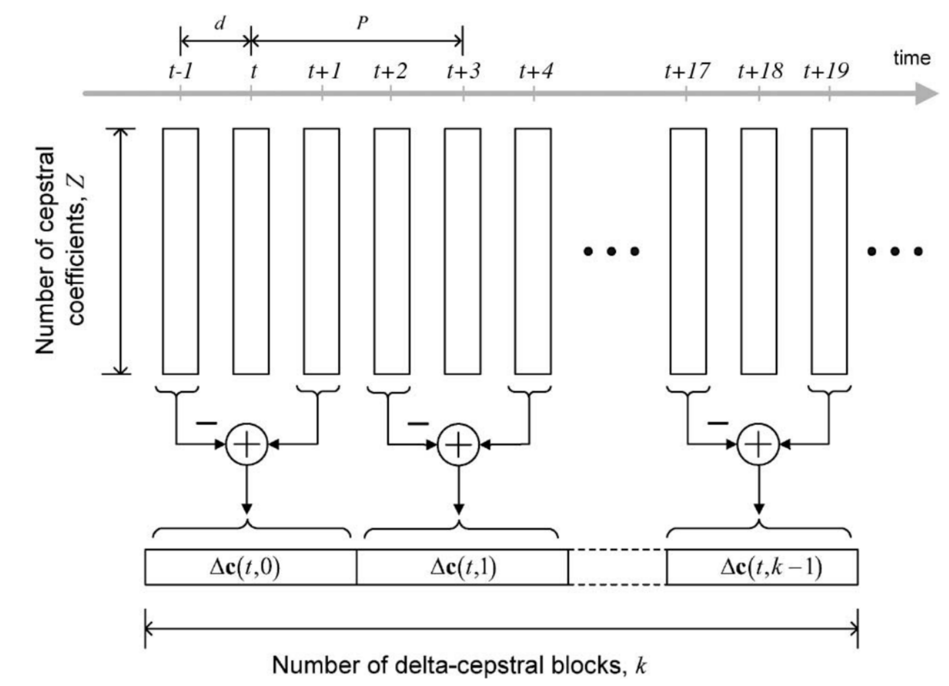
❸ **Transition to Mel Scale**: $\hat{x}_{\text{Mel}}(f) = \hat{x}(\mathcal{M}(f))$



Construction of Cepstral Coefficients

❹ **DCT of logarithm**: $x_{\text{Cep}}[i] = \text{DCT}(\log \hat{x}_{\text{Mel}}(f))[i]$

❺ **SDC**: $x^{k}_{\text{SDC}}[i] = x^{k+1}_{\text{Cep}}[i] - x^{k-1}_{\text{Cep}}[i]$



Shifted Delta Cepstral Coefficients

- 7-vectors cepstral coefficient for each 10 ms segment
- Use 3 adjacent cepstral coefficients to construct a 7-vector SDC coefficient
- For each 210ms sample ⇒ 49 SDC coefficients
- $46 \times 376 = 17296$ data samples $x^{(i)}_{\ell}$

Gaussian Mixture Model

Density Estimation

- Use $x^{(i)}_{\ell}$ to estimate mixture of N Gaussian densities

$$p_{\ell}(x) = \Sigma_{k=1}^{N} w^{i}_{\ell} P(x; \mu^{i}_{\ell}, \Sigma^{i}_{\ell}), \quad x \in \mathbb{R}^{49}$$

$$P(x; \mu^{i}_{\ell}, \Sigma^{i}_{\ell}) = \frac{1}{(2\pi|\Sigma^{i}_{\ell}|)^{49/2}} e^{-\frac{1}{2}(x-\mu^{i}_{\ell})^{T}[\Sigma^{i}_{\ell}]^{-1}(x-\mu^{i}_{\ell})}$$

$$\mu^{(i)}_{\ell} \in \mathbb{R}^{49}, \Sigma^{(i)}_{\ell} \in \mathbb{R}^{49\times 49}$$

- $\Sigma^{(i)}_{\ell}$- either `full` or `diag`.
- Training using Python's `sklearn` toolkit.

Classification

- Given 10s sample, construct

$$x^{i}, \quad i = 1, \dots, 46$$

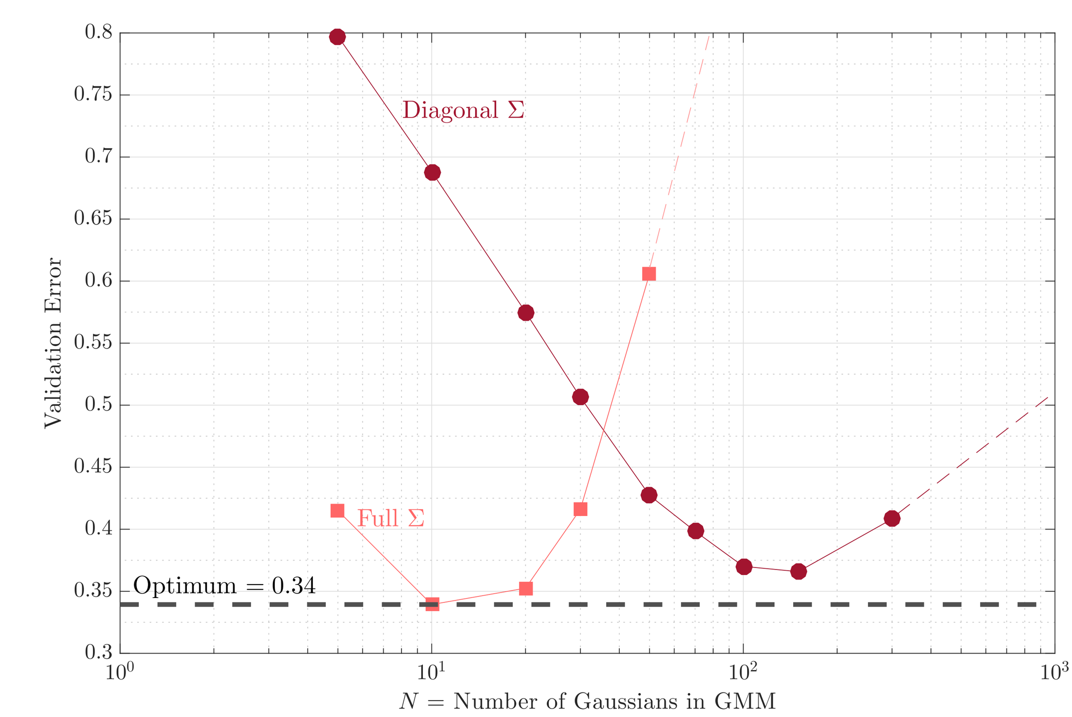
SDC vectors for each 210 ms segment.

- MLE estimator

$$\hat{\ell} = \arg\max_{\ell \in \mathcal{L}} \prod_{i=1}^{46} p_{\ell}(x_i)$$

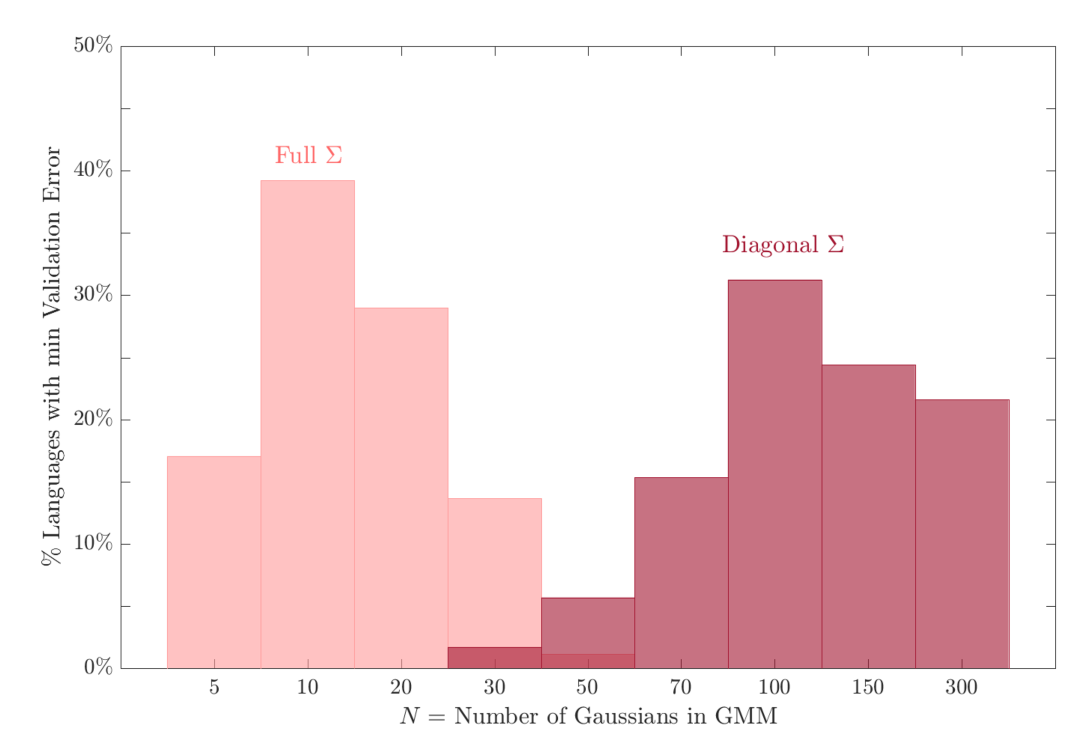
Results

- Model trained with full and diagonal covariances
- Optimal for full covariance at $N = 10$



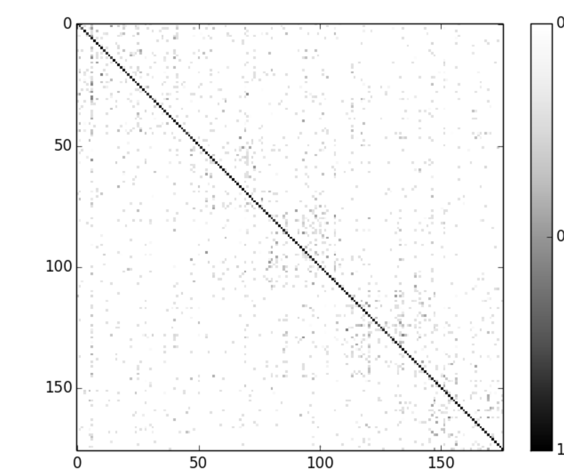
Validation Error as Function of N

- Optimal N actually varies significantly from one language to another

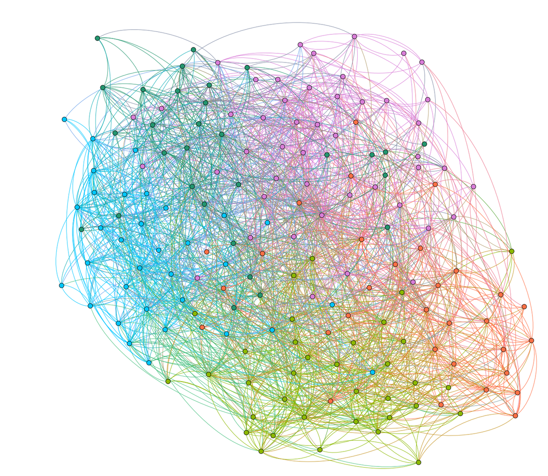


Distribution of Optimal N over Languages

- Error is non uniformly distributed but tends to cluster
- Analysis of confusion matrix graph shows communities



Confusion Matrix after Reordering



Confusion Matrix Graph with Highlighted Communities

Discussion

- Test error of 0.34 with $N = 10$ and full covariances, on a dataset with 176 (balanced) classes
- Increasing N (to some extent) decreases the validation error, which shows some consistency between the model and the data
- Relatively high variability in the accuracy depending on the language

Future Work

- Allow variability in N per language for model fitting
- Introduce a measure for the quality of the GMM fit
- Train the algorithm on a larger dataset
- Predict how well would the classifier fit the sample (confidence intervals)

Bibliography

[1] Li, Haizhou, Bin Ma, and Kong Aik Lee. 'Spoken language recognition: from fundamentals to practice.' Proceedings of the IEEE 101.5 (2013): 1136-1159.

[2] Torres-Carrasquillo, Pedro A., et al. 'Approaches to language identification using Gaussian mixture models and shifted delta cepstral features.' INTERSPEECH. 2002.