Applications of Statistical Machine learning models in Indic Language Speech and Text processing tasks

Rashaad Baig - 210101085   Sreehari C - 210101101   Dhanesh V - 210101117   Ketan Singh - 210101118

Abstract

This project, titled 'Applications of Statistical Machine learning models in Indic Language Speech and Text Processing Tasks', focuses on three core objectives. It employs Gaussian Mixture Models (GMM) for Speaker and Language Identification in diverse Indian languages. Isolated Speech Recognition is tackled using Hidden Markov Models (HMM), deciphering patterns in spoken language. Additionally, Part-of-Speech (POS) tagging is implemented for sentence analysis across various linguistic contexts. The project aims to get a deep understanding of statistical machine-learning methods and their various use cases pertaining to datasets and issues in the Indian Context.

1. Problem Motivation

The proposed project's applications span voice assistants, call centers, and security systems for Speaker and Language Identification using GMM in Indic Languages. Isolated Speech Detection using HMM aids in speech recognition and audio indexing, contributing to streamlined single-command processing in diverse IoT-driven devices. Additionally, POS tagging of Indic sentences supports machine translation, sentiment analysis, and information retrieval, showcasing the project's versatility in addressing linguistic challenges in an Indic setting through ML techniques.

2. Intended Experiments and Methods

The following are the intended experiments that we have chosen to perform.

2.1. Speaker and Language Identification from Voice

2.1.1. Dataset

For the speaker identification task, we plan to generate dataset on our own consisting of the data of four of us. For Language Identification task, we plan to use a subset (Hindi,Telugu and Malayalam) of Microsoft Research Speech Corpus

2.1.2. Intended model

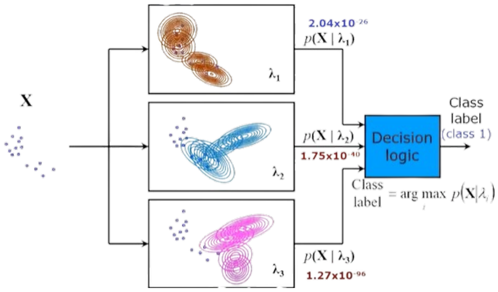
X
λ_1
λ_2
λ_3
2.04x10^{-26}
$p(\mathbf{X} \mid \lambda_1)$
$p(\mathbf{X} \mid \lambda_2)$
1.75x10^{-40}
$p(\mathbf{X} \mid \lambda_3)$
1.27x10^{-96}
Decision logic
Class label
(class 1)
Class label $= \arg\max_i p(\mathbf{X}|\lambda_i)$

Figure 1. GMM for each class of speakers to be identified

- Gaussian mixture models can be used to model broad acoustic classes in the voice of a speaker. Each speaker's acoustic features are modelled as mixture of gaussians and a given test voice data will be mapped to the GMM with maximum posterior probability of observed acoustic sequence.
- The language Identification task is a **Classification** problem in which each language to be identified is modelled as a GMM and a similar process is carried out.

2.2. Speech recognition of Isolated Indic Language Words

2.2.1. Dataset

We have planned to tentatively use Tamil Digits Dataset for this task, which contains utterances of Tamil digits from 0-10.

2.2.2. Intended Model

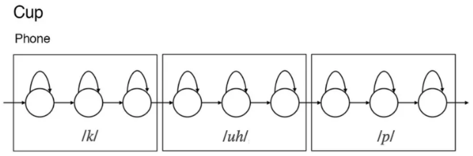

Figure 2. HMM with states as phonemes of a word *cup*

- After pre-processing, we utilize the obtained feature vector to construct a **Markov chain** required for the

HMM,because obtained feature vector can be thought of as a time series data with each dependent on the previous ones. **HMM**s naturally capture these temporal dependencies through their probabilistic framework, making them well-suited for modeling sequential data like speech.

- In this chain, every state corresponds to a ”phoneme,” which represents a fundamental unit of sound, along with a probability distribution indicating the likelihood of generating the feature vector from that state. Moreover, we incorporate a transition probability matrix, detailing the probabilities of transitioning between different states in the chain.

- The parameters of the HMM are calculated from the training data using **Maximum Likelihood Estimation** and **Vieterbi algorithm**.

2.3. Part-of-Speech Tagging of Sentence

2.3.1. DATASET

We plan to use the dataset with tagged POS from Hindi (Original) sections of the Universal Dependencies corpus.

2.3.2. INTENDED MODEL

- Here also we plan to use **HMM** sequential POS tagging, where each state represents the tag and outputs represent the words.

- A POS tagger based on HMM assigns the best tag to a word by calculating the forward andbackward probabilities of tags along with the sequence provided as an input.

- The transition probability from one state (Tag) to another is computed by calculating the respective frequency count of tags in the corpus in training phase.

3. Tentative Timeline

- **Week 1 of March:** Setup and gather the datasets, Studying necessary theory, dataset collection and preprocessing.

- **Week 2-3 of March:** Model Implementation, Training and evaluation.

- **22nd March Mid-term Report:** Provide a well-written, concise documentation of progress made so far, along with results found in the initial parts of the project.

- **Week 4 of March:** POS tagging part modelling and testing.

- **Week 1 of April:** Model Comparisions with existing models, optimisations.

- **Week 2 of April:** Documentation, Finalization of evaluation results and final report generation.

4. Discussions

- We plan to execute preprocessing like resampling the audio signals to a standardized 44.1 KHz sampling rate, employing **Voice Activity detection** (VAD)(if sample is noisy).

- We plan to explore various methods to choose number of clusters in GMM like **Silhouette score** checks how much the clusters are compact and well separated and **Bayesian information criterion** (BIC), which penalizes models with big number of clusters to avoid overfitting.

- We plan to experiment the effect of various **Covariance matrix** types for Gaussian Mixture models like Full, Diagonal, Tied etc. to determine the kind of clustering that would be achieved.

- We plan to contrast various feature extraction methods for speech like **Mel Frequency Cepstral Coefficients** (MFCCs), alongside their derivatives such as **MFCC Deltas**,**Accelerated MFCCs** and **Linear Prediction Cepstral Coefficients** (LPCCs) by comparing the performance of our models. These Cepstral Coefficients describe the overall features of the spectral envelope, will be the input for all our models using speech data in the pipeline.

- We plan to discuss the effect of number of **Hidden State** in the performance of Hidden Markov Model.

- We also plan to use **Ensemble Learning** to try to improve the predictions of a single model.

- For comparison metrics, We will use Accuracy, **Recognition Rate** for isolated word identification and **False Acceptance Rate** (FAR) and **False Rejection Rate** (FRR) to gauge language identification accuracy.

5. Additional Enhancements (If time permits)

- In practical scenarios, speech data often involves multiple speakers, necessitating the exploration of speaker segmentation (Clustering the feature vectors obtained after pre-processing into different speakers) using unsupervised learning methods like the K-Means clustering algorithm and **Hierarchical Density-Based Spatial Clustering of Applications with Noise** (HDBSCAN).

- We plan to compare and contrast with various other Classical Machine learning algorithms for classifying isolated speech like **Random Forest**.

References

Athiyaa, N. and Jacob, D. G. Spoken language identification system using mfcc features and gaussian mixture model for tamil and telugu languages. *International Research Journal of Engineering and Technology (IRJET)*, 06, 2019.

Bărbulescu, A. and Morariu, D. Part of speech tagging using hidden markov models. *International Journal of Advanced Statistics and ITC for Economics and Life Sciences*, 10:31–42, 12 2020. doi: 10.2478/ijasitels-2020-0005.

Gales, M., Young, S., et al. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3), 2008.

G.SUVARNA, K., RAJU, K.A.PRASAD, CPVNJ, D., and P.Satheesh. Speaker recognition using gmm. *International Journal of Engineering Science and Technology*, 2, 06 2010.

Irianto, A. B. P. Voice recognition using k-means clustering based on hidden markov model. In *2019 2nd International Conference on Applied Engineering (ICAE)*. IEEE, 2019. doi: 10.1109/ICAE47758.2019.9221696.

Joshi, N., Darbari, H., and Mathur, I. Hmm based pos tagger for hindi. volume 3, 09 2013. ISBN 9781921987007. doi: 10.5121/csit.2013.3639.

Reynolds, D. *Gaussian Mixture Models*. Springer US, Boston, MA, 2009. ISBN 978-0-387-73003-5. doi: 10.1007/978-0-387-73003-5_196.

Reynolds, D. and Rose, R. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3 (1), 1995. doi: 10.1109/89.365379.

Soehardinata, J. A Brief Introduction to Gaussian Mixture Model (GMM) Clustering in Machine Learning — joey.soehardinata. https://shorturl.at/huU78.