SPEAKER VERIFICATION USING GAUSSIAN MIXTURE MODEL

Manik Sharma (2021336)

Abstract

This research paper delves into speaker verification using Gaussian Mixture Models (GMMs) on the given dataset. Speaker verification is a crucial task in voice recognition systems and security applications. The study comprehensively investigates the application of GMMs for verifying speaker identities on the dataset, an extensive collection of audio recordings. It evaluates the performance of the GMM-based speaker verification system using standard metrics such as Equal Error Rate (EER). The paper also explores various aspects of feature extraction using MFCCs, pitch and PLP, considering the impact of different feature sets on verification accuracy. The findings of this research offer valuable insights into the efficacy of GMMs for speaker verification. They can have practical implications in voice biometrics, access control, and authentication systems.

1 Introduction

Speaker signals carry the intended message and additional details like the speaker's identity, room characteristics, and the type of handset used. Speaker recognition aims to extract information about the speaker, prioritizing speaker identity over the content of the message.

A speaker verification system aims to determine whether a given voice sample corresponds to a specific known reference speaker. This process entails comparing the stored model of the claimed speaker's utterance with the voice sample provided by the claimant. If the match exceeds a predefined threshold, the identity claim is accepted. Speaker verification focuses on establishing the speaker's identity rather than analyzing the message content.

Speaker recognition can be classified into identification and verification. Speaker identification is determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker.[4] The system we will describe is classified as a text-independent speaker identification system since it identifies the person who speaks regardless of what people say. This method requires the speaker to provide utterances of keywords or sentences, the same text used for training and testing.[5]

2 Literature Survey

GMM UBM is a speaker verification technique that uses a Gaussian mixture model (GMM) to represent the speaker's voice and a universal background model (UBM) to represent the general characteristics of speech. The UBM is a large GMM trained on a diverse set of speakers and speech data. The speaker's GMM is obtained by adapting the UBM using the speaker's enrollment data, which is a set of speech samples from the speaker. The adaptation process usually involves updating the mean vectors of the UBM using the maximum a posteriori (MAP) criterion. The speaker verification is done by computing the likelihood ratio of the test speech sample given the speaker's GMM and the UBM. If the ratio is above a certain threshold, the speaker is accepted; otherwise, the speaker is rejected.

You can add the following paragraph to your literature survey to introduce GMM UBM:

One of the most widely used speaker verification

techniques is the Gaussian mixture model/universal background model (GMM-UBM) approach1. This technique models the speaker's voice using a GMM, which is a weighted sum of multivariate Gaussian distributions. Each Gaussian component captures a specific acoustic characteristic of the speaker's voice, such as the spectral shape of a vowel or a consonant. The GMM is adapted from a universal background model (UBM), which is a large GMM trained on a diverse set of speakers and speech data. The UBM represents the general characteristics of speech and serves as a reference model for speaker verification. The adaptation process usually involves updating the mean vectors of the UBM using the maximum a posteriori (MAP) criterion, which preserves the speaker-independent information while incorporating the speaker-specific information. The speaker verification is done by computing the likelihood ratio of the test speech sample given the speaker's GMM and the UBM. If the ratio is above a certain threshold, the speaker is accepted; otherwise, the speaker is rejected. The GMM-UBM technique has been shown to be effective and robust for various speaker verification tasks, especially in text-independent scenarios.[6]

3 DataSet

This dataset contains speeches of five prominent leaders namely; Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Tacher and Nelson Mandela which also represents the folder names. Each audio in the folder is a one-second 16000 sample rate PCM encoded. A folder called background noise contains audios that are not speeches but can be found inside and around the speaker environment e.g audience laughing or clapping. Voice Acitivity Detection(VAD) was applied to identify and extract speech segments, effectively removing silent files from the dataset by removing audio with energy less than 25 percentile of entire dataset. There were many files with only clapping sound in it so we also removed these files by applying wiener filter and thresholding for energy of each file. (some low energy speech files were lost in this step) but that will not affect our model prediction much. Our Final Dataset consist of 50% clean speech of all the speakers + 50% noisy speech of all the speakers with randomly added noise from background noise folder + audience noise + outdoor noises each divided into 1 sec segments.

4 Feature Extraction

Feature extraction plays a crucial role in the speaker verification system, significantly influencing overall system performance. In the proposed method, features are derived using MFCC, and including pitch frequency further enhances the development of a text-independent speaker verification system.

Commonly employed features for speaker verification encompass MFCC, phase information, and spectral features, among others. Integrating MFCC with additional features contributes to heightened system efficiency. Evaluation metrics for speaker verification systems typically include the Equal Error Rate (EER) and accuracy in scenarios with degraded conditions, such as noisy environments; vowel onset points can serve as valuable features. For datasets with limited durations, utilizing expected log-likelihood with GMM-UBM becomes instrumental in the speaker verification process.

Techniques such as Cepstral Mean Subtraction (CMS) and noise masking are often applied to enhance the features' robustness. The computation of Mel-Frequency Cepstral Coefficients (MFCCs) are derived from the signal's power spectrum. Typically, 13 coefficients are calculated to represent the spectral characteristics of the signal. To capture dynamic information over time, the first-order and second-order derivatives of the MFCCs, known as delta and delta-delta coefficients, are computed. This results in a total of 39 features (13 MFCCs + 13 delta coefficients + 13 delta-delta coefficients). Furthermore, for noise robustness, noise masking is employed. This technique involves identifying and suppressing noise components in the signal, ensuring that the extracted features are less susceptible to the influence of environmental noise. Noise Masking: adding white/pink noise in a dataset to suppress unwanted noise. Cepstral Mean Subtraction: Subtracting each MFCC

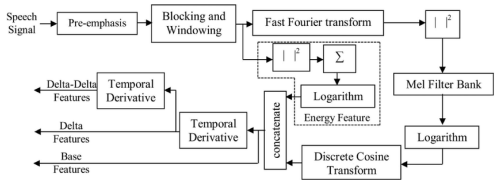


Figure 1: Block diagram of MFCC

channel by its mean to remove unwanted convolutive noise.

5 Methodology

The distinctive characteristics inherent in speech signals, which vary from one speaker to another, play a pivotal role in our system, given our primary objective of authenticating the true speaker. The success of our verification decision hinges on our ability to extract pertinent information from the speech signals effectively.

Various features are employed in speaker recognition systems, including Frequency Band Analysis, Formant Frequencies, Pitch Contours, and Harmonic Features, among others.

Standard methods for extracting speech features encompass MFCC, LPC, LPCC, RCC, LFCC, EFCC, CFCC, and phase information. Among these, MFCC is a widely utilized method acknowledged for its effectiveness in speech and speaker recognition tasks. The MFCC extraction method leverages knowledge about human auditory perception by applying the Discrete Cosine Transform (DCT) to the logarithm of the short-term energy spectrum, which is transformed using a nonlinear Mel-frequency scale. The Mel-frequency scale introduces a frequency warping to approximate the unequal sensitivity of human hearing across different frequencies.

This paper employs MFCC and pitch as the features of the speaker system. The block diagram illustrating the computation of MFCC using DCT is depicted in Figure 2.[2] The process involves segmenting the speech signal into blocks using overlapping smooth Hamming windows, followed by taking the FFT of the windowed signal. Subsequently, the mel-scaled filter bank is computed, and the logarithm of the mel-scaled filter bank is determined. The final step involves applying the Discrete Cosine Transform to the mel-scaled filter bank to calculate the MFCC.

An additional extracted feature is the pitch frequency of the speech signal, which represents the fundamental frequency of vocal cord vibrations. The pitch period of the signal is determined by the number of samples, after which the waveform repeats itself. Taking the inverse of this pitch period yields the pitch frequency of the signal. Several methods are employed to calculate pitch frequency, including autocorrelation and average magnitude difference function (AMDF).

5.1 Statistical Modeling

The core of the speaker verification decision is the likelihood ratio. Say that we want to determine if speech sample Y (from the verification set) was spoken by S. Then, the verification task is a basic hypothesis testing: H0: Y is from speaker S H1: Y is not from speaker S, The test to decide whether to accept H0 or not is the Likelihood Ratio (LR):

$$LR = \frac{p(Y \mid H_0)}{p(Y \mid H_1)}.$$

If the Likelihood ratio is greater than the threshold , we accept H0: otherwise we accept H1. If we talk in terms of logs, then the log-likelihood ration is simply the difference between the logs of the 2 probability density functions:

$$\log(LR) = \log(p(Y \mid H_0)) - \log(p(Y \mid H_1)).$$

We have a speaker to test for H0,and we can build a model, say

$$\lambda_{hyp}.$$

,being for example a Gaussian Distribution of the features extracted. However, we do not have an alternative model for H1 . We must compute what is called a "Background Model", which would be a Gaussian Model

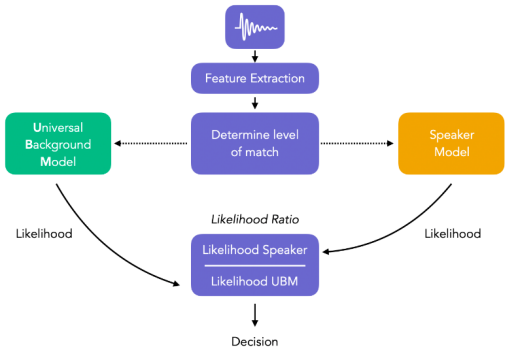
$$\lambda_{\overline{hyp}}$$

. There are 2 options for the background model: either consider the closed set of other speakers and compute

$$p(X \mid \lambda_{\overline{hyp}}) = f(p(X \mid \lambda_1), ..., p(X \mid \lambda_N)).$$

, where f is an aggregative function like the mean or the max. It however requires a model per alternative hypothesis, i.e. per speaker or consider a pool of several different speakers to train a single model, called the Universal Background Model (UBM).

The Pipeline can be represented as such:



5.2 Universal Background Model : Development

A UBM is a high-order Gaussian Mixture Model (usually 512 to 2048 mixtures with 24 dimensions) trained on a large quantity of speech, from a wide population. This step is used to learn speaker-independent distribution of features, used in the alternative hypothesis in the likelihood ratio.

For a D-dimensional feature vector x , the mixture density is:

$$P(x \mid \lambda) = \sum_{k=1}^{M} w_k \times g(x \mid \mu_k, \Sigma_k).$$

Where:

c is a D-dimensional feature vector

$w_k, k = 1, 2, \ldots, M$ are the mixture weights s.t. they sum to 1

$\mu_k, k = 1, 2, \ldots, M$ are the mean of each Gaussian

$\Sigma_k, k = 1, 2, \ldots, M$ are the covariance of each Gaussian

$g(\mu_k, \Sigma_k)$ are the Gaussian densities such that:

$$g(\mu_k, \Sigma_k) = (2\pi)^{-D/2} |\Sigma_k|^{1/2} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

We typically use a diagonal covariance-matrix rather than a full-covariance one since it is more computationally efficient and empirically works better. The GMM is trained on a collection of training vectors. The parameters of the GMM are computed iteratively using Expectation-Maximization (EM) algorithm, and therefore there are no guarantees that it will converce twice to the same solution depending on the initialization. Under assumption of independent feature vectors, the log-likelihood of a model for a sequence is simply the average over all feature vectors:

$$\log p(X \mid \lambda) = \frac{1}{T} \sum_t \log p(x_t \mid \lambda)$$

5.3 Speaker Enrollment

The last step before the verification is to perform the speaker enrollment. The aim is still to also train one Gaussian Mixture Model on the extracted features for each speaker, thus resulting in 5 models.

There are 2 approaches we used to model the speakers: 1. Trained a lower dimensional GMM (10-20) depending on the amount of enrollment data that we have adapt the UBM GMM to the speaker model using Maximum a Posteriori Adaptation (MAP), usually the approach selected 2.In MAP, we simply start the EM algorithm with the parameters learned by the UBM. Through this step, we only adapt the mean, and not the covariance, since updating the covariance does not improve the performance. For the mean to update, we perform a maximum a posteriori adaptation :

$$\mu_k^M AP = \alpha_k \mu_k + (1 - \alpha_k) \mu_k^U BM$$

$\alpha_k = \frac{n_k}{n_k+\tau_k}$ is the mean adaptation coefficient.
n_k is the count for the adaptation data
τ_k is the relevance factor, between 8 and 32

5.4 Gaussian Mixture Model (GMM)

The speaker model is created using the Gaussian mixture model (GMM). For making the model, first, a discriminant function gi(x) = log f(x|i) is developed where the density f(x|i) is a linear combination or mixture of L multivariate Gaussian densities, which is given by [2]

$$f(x|i) = \sum_{j=1}^{L} w_j b_j(x) \quad (1)$$

Figure 2: Equation of Multivariate Gaussian Density

The paper autocorrelation method is used to find pitch where we are the scalar weights applied to each Gaussian probability density function given by bj(x). Its mean vector and covariance matrix can specify this function. The sum of the weights wj equals 1. The GMM classifier functions as a Bayesian discriminant, with speaker models gi(x) represented as Gaussian, inherently multimodel mixture densities. To construct the GMM, the number of mixtures (L) is determined, and a maximum likelihood formulation is employed for establishing the GMM parameters. The iterative Expectation-Maximization (EM) algorithm is pivotal in attaining the maximum likelihood ratio. In speaker recognition, an ensemble of test feature vectors x1,x2....,xq are utilized.

The speaker's identification is based on selecting the model that maximizes the score. The test score is compared against a predetermined threshold in open-set speaker verification scenarios. A test is conducted exclusively with a similar speaker to compute a likelihood ratio for speaker verification.

5.5 Universal Background Model (UBM)

A Universal Background Model (UBM) is a high-order Gaussian Mixture Model, typically comprising 512 to 2048 mixtures with a dimensionality of 24. It is trained on an extensive dataset of speech samples representing a diverse population. This training process aims to acquire a speaker-independent distribution of features, serving as the foundation for the alternative hypothesis in the likelihood ratio.

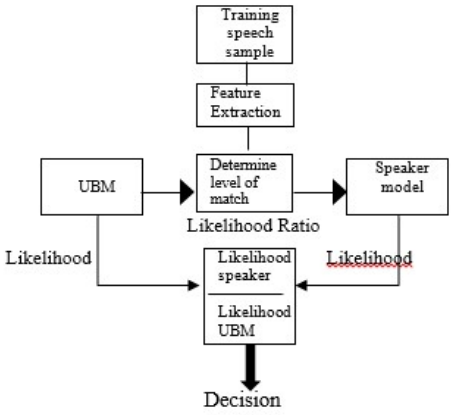


Figure 3: Pipeline of UBM model

5.6 Maximum Likelihood Estimation

MLEs are the parameters' elements that increase the probability of the observed items. Parameter estimation for GMM using maximum likelihood, λ

Denotes an initial model for ML. The mean and variance are estimated from the known data to maximize the likelihood function[6][7]

$\text{Ln}[\text{L}(\lambda|x1, x2, x3....xn)] = \pi_{i=1}^{n} Lnf(xi|\lambda)$[8]

Working with the natural logarithm of the likelihood function is more convenient. The mixture model represents probabilistic data belonging to the distribution of the mixture model. Each element of the mixture's components is a Gaussian distribution with its unique parameters and corresponding variance variables.

5.7 Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm is an iterative method commonly used for finding maxi-

mum likelihood or maximum a posteriori estimates of parameters in statistical models, particularly when dealing with incomplete or latent data. It consists of two main steps: the Expectation (E) step and the Maximization (M) step. The EM algorithm serves as a means to attain the maximum probability, which is especially useful when dealing with incomplete or unexpected data. This iterative method is repeatedly employed to identify the maximum potential task.

[1.]First, initialize the λ parameters for some random values.

[2.]For each possible value of Z, compute the probability by given λ.

[3.]Then, use only calculated Z values to estimate λ parameters better.

[4.]Repeat steps 2 and 3 until convergence.[6]

Symbols	Description
λ	likelihood of GMM model
Z	the probability of each possible value
Ln L(x)	Natural logarithm of likelihood function

Table 1: Symbols and their description

6 Analysis and Results

The table below summarizes the Equal Error Rate (EER) values for the speaker verification experiments using different configurations.

1.

UBM Components	**GMM Components.**	**Train Data EER**	**Test Data EER**
40	16	0.05	0.22
34	8	0.07	0.23
16	8	0.10	0.28

UBM using MAP Components	**Train Data EER**	**Test Data EER**
128	0.0698	0.0826
64	0.0812	0.0955
32	0.9231	0.1049
16	0.1021	0.1126
8	0.1219	0.1371

References

[1] Kerlos Atia Abdalmalak & , Ascension Gallardo-Antolin (2016) *Enhancement of a Text-Independent Speaker Verification System by using Feature Combination and Parallel-Structure Classifiers* Spain/Springer–Link.

[2] Shilpa S. Jagtap & D.G.Bhalke (2015) Speaker Verification Using Gaussian Mixture Model *International Conference on Pervasive Computing (ICPC).*

[3] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, vol. 17, no. 1-2, pp. 91-108, 1995.

[4] Saeidi, Rahim. "Advances in Front-end and Back-end for Speaker Recognition." nation: a feature-based approach 13.5 (2011): 58-71.

[5] A. Mansour and Z. Lachiri, "SVM-based Emotional Speaker Recognition using MFCC-SDC Features", International Journal of Advanced Computer Science and Applications, vol. 8, no. 4, 2017.

[6] D. A. Reynolds, "An overview of automatic speaker recognition technology," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), vol. 4, pp. 4072–4075, Orlando, Fla, USA, May 2002.